Modelling multimodal language processing

Cover design: Heidi Caunce Berry Design, www.heidicaunceberry.co.uk

ISBN: 978-90-76203-71-3

Printed and bound by Ipskamp Drukkers, Nijmegen.

Modelling multimodal language processing

Proefschrift

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de rector magnificus,

volgens besluit van het college van decanen

in het openbaar te verdedigen op maandag 7 december 2015

om 12.30 uur precies

door

Alastair Charles Smith

geboren op 11 september 1983

te Birmingham, Verenigd Koninkrijk

Promotoren

Prof. dr. A.S. Meyer

Prof. dr. P. Monaghan (Lancaster University, Verenigd Koninkrijk)

Copromotor

Dr. F. Huettig (Max Planck Institute for Psycholinguistics)

Manuscriptcommissie

Prof. dr. A.P.A. Roelofs

Prof. dr. C.N.L. Olivers (Vrije Universiteit Amsterdam)

Dr. S.L. Frank

The research reported in this thesis was supported by the Max-Planck-Gesellschaft zur Förderung der Wissenschaften, München, Germany.

Contents

Chapter 1

General Introduction

A fundamental property of language is its ability to connect information across modalities. For example on hearing the spoken word "apple" or viewing the written word *apple* we are able to rapidly generate knowledge of an apple's visual properties (e.g. its shape, its colour), knowledge of its olfactory properties (e.g. how an apple may smell), knowledge of its gustatory properties (e.g. how an apple may taste), knowledge of its tactile properties (e.g. how an apple may feel to touch) and knowledge of its functional semantic properties (e.g. it is edible, it is healthy) among many others. Conversely, on viewing an apple, smelling the scent of an apple, experiencing the taste of an apple, or recalling from memory a healthy, edible, fruit we are able to generate rapidly knowledge of the spoken or written form of the word *apple*. The speed and ease with which the cognitive system performs these complex operations has been noted and intrigued thinkers for centuries:

"No sooner do we hear the words of a familiar language pronounced in our ears but the ideas corresponding thereto present themselves to our minds: in the very same instant the sound and the meaning enter the understanding: so closely are they united that it is not in our power to keep out the one except we exclude the other also. We even act in all aspects as if we heard the very thoughts themselves." G. Berkeley, An Essay Towards a New Theory of Vision, Dublin, 1709.

This short extract highlights questions that are fundamental to our understanding of language processing, thus fundamental to the field of psycholinguistics and that lie at the heart of the investigations detailed in this thesis. A complex yet effortless process is initiated when an individual is exposed to the bundle of auditory properties carried in the speech signal or visual properties carried in the visual signal that constitute the event of hearing or viewing a word (e.g. 'apple'). But, what is the nature of the information activated, what is the time

course of activation, what architecture is able to support this process and do these properties of the system vary across human populations?

Assertions made regarding such properties of the language processing system unsurprisingly connect with a number of major debates within the field of psycholinguistics and cognitive science more broadly. For example debates regarding the time course of multimodal interaction, at what stage of processing can information within one modality influence processing in another. In 1705 Berkeley asserted that "…in the very same instant the sound and the meaning enter the understanding…", how is it that words embedded within noisy auditory speech signals can activate such rich representations of items' meanings so rapidly. Berkeley's assertion of immediate activation aligns with parallel interactive models of mutlimodal processing. Such models posit that knowledge from multiple, distinct information sources are integrated immediately and in parallel (e.g. Rumelhart & McClelland, 1986). Therefore, in the case of speech processing, pre-activated semantic or visual properties of objects in the local environment can immediately influence processing of the auditory signal (e.g. McClelland and Elman, 1986; Gaskell & Marslen-Wilson, 1997; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995; Dilkina et al, 2009; Ueno et al., 2011; Pulvermuller et al., 2009; Rogers et al., 2004). This contrasts with modular cascaded (e.g. McClelland, 1979) or serial models (e.g. Chomsky, 1965; Fodor, 1983) which argue that information outside of a given modality cannot influence processing within the modality either at early stages of processing (cascaded) or until processing within the given modality is complete (serial). Therefore, in the case of speech, either initial (e.g. Caramazza, 1997; Dell, 1986; McQueen, Dahan & Cutler, 2003; Huettig & McQueen, 2007) or for all (see Levelt, 1999; Chater & Manning, 2006; Freiderici, 2002; Morton, 1969; Shallice, 1988; Cutler & Norris, 1979; Norris, McQueen, & Cutler, 2000) processing of the speech signal information active in semantic or visual domains has no influence on processing. Within this thesis I use computational models to generate predictions of the neural and behavioural consequences of a language system supported by a parallel interactive architecture. These predictions are then evaluated in relation to existing (Chapters 2, 3, 5 & 6) and novel (Chapter 4) data sets.

Berkeley also states that "…so closely are they [sound and meaning] united that it is not in our power to keep out the one except we exclude the other also." This statement is most consistent with interactive models of cognition in which representations emerge from a continuous process in which multiple forms of information interact in parallel (Gaskell & Marslen-Wilson, 1997; Rogers et al., 2004; Dilkina, McClelland & Plaut, 2008; Ueno,

Sauito, Rogers & Lambon-Ralph, 2011). This contrasts with modular systems in which stored properties of an item are accessed via a sequence of discrete modality specific processing stages (Levelt, 1999; Chater & Manning, 2006; Freiderici, 2002; Morton, 1969; Shallice, 1988; Cutler & Norris, 1979; Norris, McQueen, & Cutler, 2000). Further, this statement not only implies an interactive speech processing system, but also connects with another significant debate within cognitive science regarding the structure of representations supporting multimodal cognitive processing. Specifically is language processing, and semantic processing more broadly, dependant on the development of amodal representations. Interactions between modalities can either be supported via direct connections between individual modalities (e.g. Wernicke and Meynert, see Eggert 1977; Barselou, 1999; Wilson, 2002; Gibbs, 2006), or via shared resources that facilitate interaction between various distinct modalities (see Rogers et al. 2004; Plaut, 2002; Lambon-Ralph & Patterson, 2008; Dilkina, McClelland & Plaut, 2008). This second perspective allows for the development of amodal representations, that encode knowledge of an item independent of a single modality, thus within such representations knowledge of the item's sound, for example, cannot be distinguished from knowledge of its meaning. Within this thesis I implement computationally an interactive model of multimodal language processing in which individual modality specific information processing streams are connected via a central shared resource. The structure of representations developed within such an architecture are then examined (Chapter 3) and their behavioural consequences evaluated against existing (Chapters 2 and 3) and novel data sets (Chapter 4).

Finally, Berkeley asserts "We even act in all aspects as if we heard the very thoughts themselves." This statement acknowledges a critical property of language that is needed for communication to take place. In order for the given word to generate the affect the sender had intended the relationship between a given word and its meaning must possess properties common to both the individual sending the message and the individual that receives the message. This begs the question, how are relationships between word and meaning formed, to what extent is this process shared between individuals such that it generates relationships between word and meaning that are common across individuals. How words gain their meaning or in other words how language is grounded or connected to the world in which it functions is often referred to as the 'symbol grounding problem', (Harnad, 1990). A potential solution to this problem is that language is grounded in our perceptual experience of the shared world that surrounds us, embodiment theorists argue that the meaning of words is

grounded in the "representational codes specific to our perceptual systems" (Prinz, 2002). However, as the earlier paragraph discussed, models grounded in such a way differ in the level to which language processing is dependent on representations that are amodal and abstracted away from modality specific perceptual systems (see Lambon-Ralph & Patterson, 2008; Barsalou, 2008; for review). Grounding language in this way offers explanation for how communication can function. Language facilitates effective communication between individuals as they possess common connections between the perceptual properties that constitute hearing or seeing the spoken or written form of a word (i.e. apple) and the perceptual properties that constitute its meaning (i.e. green, circular, sweet smelling, etc…). This is due to similarities in the processes that develop these connections and similarities in the learning environment (shared properties of the physical world). Within this thesis I capture in computational models the process through which a system exposed to a multimodal learning environment establishes such connections across modalities. Further as computational models they offer an explicit description of an architecture able to support such connections and as models of learning they describe the representations that emerge from this interaction (Chapters 3 & 6).

Although as Berkeley asserts, we are able to communicate ideas and thoughts effectively using language there is variation between individuals in the structure of the language processing system. To take an extreme example the relationships between word and meaning are not always shared. For a speaker of Dutch the phonological properties of the spoken word 'pop' are likely to be strongly associated with the visual, properties of a toy doll, while for an English speaker these phonological properties are likely associated with the gustatory properties of sugary, fizzy drinks. In this example variation in the language processing system comes from differences between the two populations in the associations between representations which in turn derive from differences in the information to which the cognitive system is exposed in the learning environment. Variation in language processing across individuals may arise for other reasons also. For example, differences in the structure of the architecture supporting language processing (e.g. Wright, et al., 1997; Dehaene et al., 2010) or differences in the structure of representations (e.g. Kuhl, 2000; Majid et al., 2004; Ziegler & Goswami, 2005) which may in turn be influenced by the structure of information in the learning environment or the learning mechanism that acts upon this information. Within this thesis I develop computational models in which such parameters can be manipulated directly to examine their effects on language processing and cognition more broadly.

Although as earlier argued there must be properties common to the relationships between perceptual experiences of the world, individuals each have unique multimodal sensory experience of that world. Over the course of development the pattern of sensory information to which we are exposed for example, stimulation by visual stimuli of neural populations in the retina and auditory information stimulating neural populations in the cochlear, will be unique for each individual. Within this thesis I use computational models to examine how over the course of development the structure of information in the multimodal learning environment shapes language processing. This is performed by holding the learning mechanisms and cognitive architecture I assume to be supporting language processing fixed while varying the structure of information in the learning environment (Chapters 2, 3 and 6) or structure of representations activated by the learning environment (Chapter 5).

Another aspect of Berkeley's assertion is that our behaviour reflects the manner in which the speech signal is processed, "we even *act* in all aspects as if we had heard the very thoughts themselves". Therefore, if two populations receive the same signal yet behave differently this may expose differences in properties of the underlying system. Again, using computational models I develop models that describe explicitly the connection between cognitive processing and behaviour. These models are then used to examine the extent to which differences in behaviour observed over the course of development or between populations can be explained by differences in processing as a result of quantitative or qualitative differences in the structure of information in the learning environment or the structure of stored representations (Chapters 3, 5 & 6).

Chapters 5 and 6 of this thesis focus specifically on how variation in language processing, and cognitive processing more broadly, may arise as a consequence of qualitative or quantitative differences in exposure to literacy training. Current computational and theoretical models of spoken word (e.g., Cohort Model, Marslen-Wilson & Tyler, 1980; MERGE, Norris, McQueen, & Cutler, 2000; Shortlist B, Norris & McQueen, 2008; TRACE, McClelland & Elman, 1986) and written word (Harm & Seidenberg, 1999; Coltheart et al, 2001; Harm & Seidenberg, 2004; Perry, Ziegler, & Zorzi, 2007, 2010) processing are derived largely from studies of alphabetic literate populations. However, to be described as comprehensive models of language processing they must also take into account the sizable proportion of the human population who are not alphabetic literates. For example, approximately 16% of the world's adult population is illiterate (UNESCO Institute for Statistics, 2013), while over 1 billion individuals are literate in logographic languages and

over 500 million in alphasylabic languages. Existing empirical data suggests that exposure to literacy training affects performance on a range of cognitive tasks (Reis, Guerriero & Petersson, 2003; Kosmides, Tsapkini, Folia, Vlahou & Kiosseoglou, 2004; Reis & Castro-Caldes, 1997; Szwed, Ventura, Querido, Cohen and Dehaene, 2012; Bramao, Mendonca, Faisca, Ingvar, Petersson & Reis, 2007; Olivers, Huettig, Singh & Mishra, 2014), and further that the nature of the effects may be modulated by the structure of the orthographic system on which an individual receives training (Ziegler & Goswami, 2005; Brennan, Cao, Pedroarena-Leal, McNorgan & Booth, 2013; Cao, Khalid, Lee, Brennan, Yang, Li, Bolger & Booth 2011; Cheung, Chen, Lai, Wong & Hills, 2001; Ho & Bryant, 1997; Huang & Hanley, 1995, 1997; McBride-Chang, Bialystok, Chong & Li, 2004; Read, Yun-Fei, Hong-Yin & Bao-Qing, 1986; Shu, Peng & McBride-Chang, 2008). Further, in the case of reading the structure of the orthography on which a system is trained is likely to have major implications for how the ability to read is acquired and the cognitive mechanisms recruited for reading (e.g. Frost, Katz & Bentin, 1987; Katz & Feldman, 1981; Kiyosawa et al. 1995; Paulesu et al., 2000; Seidenberg, 2013). However, isolating the effects of literacy training or orthographic system from other co-varying linguistic or non-linguistic factors has proved difficult. Computational models however, offer a means by which specific differences in the structure of the learning environment or structure of representations can be manipulated directly and in isolation to observe their effects on processing and behaviour. This is the approach taken in chapters 5 and 6 of this thesis.

Within this thesis, using emergent computational models, I offer an explicit description of the multimodal nature of spoken and written word processing. This necessitates defining the structure of the representations that operate within and across modalities and the cognitive architecture that supports this multimodal interaction (Chapters 2, 3, 5 & 6), and thus each model represents the implementation of precise positions on each of the above mentioned debates. As emergent models they also chart the course of development providing a description of the extent to which representations and processing can be determined by the structure of the multimodal learning environment (Chapters 4 & 6). Further as computational models, they allow direct manipulation of properties of the learning environment, representations or architecture that may vary across populations in order to examine in isolation their impact on emergent representations and processing. Within this thesis I exploit this property of the models to examine how differences between populations in their exposure to literacy training may impact on language processing and cognition more broadly (Chapters

5 & 6). Throughout, models are tested on their adequacy in capturing and offering explanation for behavioural and neural effects reported in the literature, while they are also used to generate predictions that are then (Chapter 4), or can in future empirical studies be, tested against novel data sets. Together this computational modelling approach offers a means of testing multiple hypotheses relating to the representations and architecture that support the complex and dynamic process that is multimodal language processing. Without explicit implementation in computational models such questions are likely to remain intractable, lying beyond the scope of current behavioural or brain imaging methods.

Methodology

This thesis focuses specifically on the interaction of two fundamental components of the human cognitive system, language and vision. With the aim of describing explicitly the representations and cognitive architecture that supports their interaction.

In chapters 2 - 5 I focus on modelling language mediated visual attention (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995), which describes changes in the distribution of visual attention as a function of the concurrent speech signal. My motivation for modelling this feature of human behaviour was three-fold. Firstly, language mediated visual attention is a behaviour common to hearing and sighted individuals performed regularly during everyday life, for example when someone is asked to “pass the salt”, “look at the camera” or “turn left at the traffic lights”. In order to perform each of these requests the individual must integrate information carried in the speech signal with visual information extracted from the local visual environment. Therefore, constructing a computational model of the system underlying this behaviour will inform our understanding of how this common behaviour is performed.

Secondly, language mediated visual attention has proved very influential in the field of psycholiguistics (see Huettig, Rommers & Meyer, 2011 for review). It has been used for example to infer properties of stored semantic knowledge (e.g. Huettig & Altmann, 2005; Huettig, Quinlan, McDonald & Altmann, 2006; Mirman & Magnuson, 2009; Yee & Sedivy, 2006; Yee, Overton & Thompson-Schill, 2009), how visual information is extracted from the local visual environment (see Huettig, Olivers & Hartsuiker, 2011 for review), the process through which syntactic frames are constructed during spoken language production (e.g. Brown-Schmidt & Konopka, 2008; Van de Velde, Meyer & Konopka, 2014) and to infer

temporal properties of the architecture supporting spoken word comprehension (e.g. Allopenna et al., 1998; Huettig & McQueen, 2007). However, how this indirect measure of eye gaze is connected to the underlying cognitive processes such studies argue it represents are often left underspecified (Ferriera & Tanenhaus, 2007; Huettig, Olivers & Hartsuiker, 2011; Anderson et al., 2011). Therefore, the aim of developing a computational model that describes this connection explicitly intends to reduce ambiguity in the linking hypothesis and thus reduce ambiguity in interpretation of such data.

Language mediated visual attention is typically recorded using the visual world paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995) in which individuals view a visual display while hearing a spoken utterance while their eye gaze is recorded. Recording eye gaze in this way provides a rich temporal measure of behaviour over the course of the unfolding spoken word, hence exposing subtle properties of behaviour that may otherwise be lost in coarser grain behavioural measures. Due to these properties of language mediated visual attention it has been used to infer temporal properties of the information activated by the concurrent visual and auditory stimulus and by extension properties of the cognitive architecture that supports spoken word processing and the integration of information from auditory and visual processing streams (e.g. see Allopenna et al., 1998; Huettig & McQueen, 2007; Anderson et al., 2011). Further, as the paradigm in general does not require complex explicit responses to language tasks (e.g. instructions often require participants to simply look at the display while listening to the spoken stimuli) it can be used to examine behaviour both across development (Huang and Snedeker, 2009; Mani et al., 2013; Mani and Huettig, submitted) and diverse populations (Mirman et al., 2008; McMurray et al., 2010; Huettig, Singh & Mishra, 2011; Huettig & Janse, in press). Such data may potentially reveal how properties of the underlying system vary across populations or over the course of development. For these reasons such a measure is well suited to the current investigation offering a comprehensive set of data points against which models can be evaluated. Developing an explicit model of the representations and cognitive architecture supporting this behaviour necessitates a description of the representations and cognitive architecture that supports the interaction of these two distinct information processing streams (language and vision). Therefore such a model is likely to assist in achieving the broader goal of informing our understanding the system supporting multimodal language processing.

This thesis focuses on developing and testing computational models of the multimodal processes that underlie spoken and written word processing. Although, recent years have seen

a growth in empirical studies of the way in which multiple information types within the environment contribute to language processing (Bahrick, Lickliter, & Flom, 2004; Kirkham, 2010; Yu & Ballard, 2007; Hollich et al., 2000; Moore, Angelopoulos, & Bennett, 1999; Yurovsky, Boyer, Smith, & Yu, 2013), theoretical and computational models of the multimodal interaction involved in processing even single words have not kept pace (Ferriera & Tanenhaus, 2007; Huettig, Olivers & Hartsuiker, 2011; Anderson et al., 2011). Computational models are valuable as they require explicit description of the processes involved. They also offer a means of maintain tractability by providing a method through which well-motivated predictions for the consequences of contrasting hypotheses relating to properties of the underlying complex and dynamic multimodal system can be generated. Further, within computational models individual properties of the system can be probed and manipulated precisely and independently which is often not possible in behavioural or brain imaging studies yet is often necessary to tease apart competing hypotheses (e.g. Dilkina et al., 2010).

The models developed and tested within this thesis are parallel distributed neural network models derived from the connectionist and interactive activation tradition (Rumelhart, McClelland & the PDP Research Group, 1986; see Rogers & McClelland, 2014; McClelland et al., 2014 for review). Such models are composed of multiple layers of non-linear processing units connected to one another via weighted connections. They are often described as cognitively plausible as this low level structure is assumed to reflect neurons and synapses within the brain (Rogers & McClelland, 2014), however in order to maintain tractability and increase the efficiency of implementation models are not implemented at the level of individual neurons. Instead many details of the real system are abstracted away with networks designed to capture the behaviour of larger connected groups of neural populations.

The computational models used within this thesis are also emergent. This reflects the fact that their behaviour develops through applying a learning algorithm to train the network. This algorithm, over the course of development, makes small changes to the weights that connect units within the network, such that the network, over time, learns to perform a specific function defined by constraints imposed by the structure of the learning environment.

This relatively simple framework captures complex properties of neural systems which have allowed it to successfully connect neural and behavioural data while simultaneously operating at a level of abstraction that is both tractable and understandable (see McClelland &

Rogers, 2014, for review). For example, representations become patterns of activation distributed across multiple units within the system; processing is interactive and dynamic such that is evolves over time; learning and long term memory are captured through changes to connection weights; all of which is dependent on the statistical structure of the environment. These fundamental properties of neural systems, for example the interactive nature of processing both within and across modalities (see Ghazanfar & Schroeder, 2006; Lewokowicz & Ghazafar, 2009), make such systems incredibly difficult to predict. Therefore, implementation using such a computational framework offers a means of gaining traction on issues relating to such complex systems.

Emergent neural network models of this type are also useful in defining the scope for information structure within the learning environment to drive complex processing characteristics and behaviour (see McClelland et al., 2010 for review). Over the course of training such networks develop sensitivity to statistical regularities in the environment and interactions between those regularities that leads to the network deriving structure from the input. By maintaining a parsimonious architecture it become possible to examine the extent to which explanations of behaviour require additional complexity within the system (which if necessary at a later stage the framework allows the modeller to build in), or can develop due to the rich structure of information in the learning environment. This is the approach taken for investigations detailed in this thesis, I implement parsimonious architectures while controlling the structure of information in the learning environment in order to identify the properties of the input that drive emergent behaviour. Emergent models also offer a description of how behaviour and processing may vary over the course of development (see Elman, 1996; Munakata & McClelland, 2003; Elman, 2005). This description of development offers a strong test of a models adequacy and sufficiency and can be critical in understanding the functioning of the mature system (Sirois et al., 2008; Westermann & Ruh, 2012).

Neural network models of the type used within this thesis require decisions to be made regarding the level to which the input reflects the true complexity of the multimodal learning environment. Attempting to implement the full complexity of the this environment is not only a practical impossibility (for example simulating the activation of the retina for one single fixation of a visual scene would require describing the activation of approximately 120 million units representing the rods and 6 million representing the cones, before we consider temporal dimensions relating changes in stimulation over the course of a single trial and then over the course of development) but also reduces the likelihood of maintaining traction on the

factors driving processing characteristics and behavioural properties of the resulting system. In the modelling undertaken for this thesis I therefore adopt a fundamentalist approach, i.e. "the model should embody only the principles that are theorised to account for the phenomenon in focus" (Kello & Plaut, 2000). This increases the ability to isolate critical aspects of the environment, representations or architecture that affect performance of the model.

Thesis Outline

In chapters 2 - 5 I describe and examine the explanatory scope of a multimodal integration model (MIM) of language processing. The model implements the hypothesis that concurrent phonological, semantic and visual information is integrated in parallel during language processing, providing an explicit description of a cognitive architecture able to support such interaction. The computational model implements a recurrent neural network architecture derived from the hub-and-spoke models of semantic processing in which information from modality specific sources is integrated within a central connecting resource (Plaut, 2002; Rogers et al., 2004; Dilkina, McClelland & Plaut, 2008, 2010). Thus, within the model spoken word recognition and comprehension are framed in terms of multimodal constraint satisfaction (MacDonald et al., 1994; McClelland, Rumelhart & Hinton, 1986; McClelland et al., 2014). As a model of language mediated eye gaze, it models gaze as a continuous measure of the simultaneous integration of information across all modalities and provides an explicit description of the relationship between the multimodal input to the language processing system and the observed behavioural output (eye gaze). The model implements a finite set of assumptions common to existing models of language mediated visual attention (see Chapters 2 – 3) with specifically few assumptions made regarding constraints on the flow of information within the system. Chapters 2 – 5 examine the extent to which emergent properties of the interaction of this parsimonious architecture with basic structural properties of the multimodal learning environment are able to simulate and offer explanation for observed multimodal effects in language processing.

In chapter 2 I test whether the assumptions implemented within MIM were sufficient to replicate independent visual, semantic and phonological competitor (competitors are items in the visual display that share properties of the spoken word in one or more dimensions e.g. beaker and beaver are phonological competitors) effects on language mediated visual attention reported within the visual world paradigm literature both in the presence and

absence of named objects within the display (Allopenna et al., 1998; Dahan & Tanenhaus, 2005; Huettig & Altmann, 2007; Yee & Sedivy, 2006; Huettig & Altmann, 2005; Mirman & Magnuson, 2009). Tests of single modality effects in chapter 2 provide a necessary precursor before extending to model multiple interactive effects as examined in chapter 3.

Investigations detailed in chapter 3 examine the emergent representations and behaviour displayed by the model both over the course of development and in the mature system. This is then evaluated in relation to developmental (Mani & Huettig, submitted) and adult (Huettig & McQueen, 2007) data from the visual world paradigm literature providing a strong test of the MIM model's adequacy. The internal processing of the model is also probed in this chapter in order to understand the processes within MIM that generate behaviour observed in human populations.

Specifically, I examine whether the representations developed within the integrative layer of the model, as it learns to map between a variety of modalities, display amodal characteristics. Or alternatively, whether representations generated are only associated with a particular input and output pairing (Amedi et al., 2005; Lambon Ralph & Patterson, 2008; McNorgan, Reid, McRae, 2011; Rogers et al., 2004). This extends previous investigations using such integrative layers which have postulated that such integrative resources may function as an amodal representation (Dilkina et al., 2008; 2010; Plaut, 2002; Rogers et al., 2004*)*.

Few studies have examined language mediated eye gaze over the course of development, those that do have shown that sensitivity to semantic and phonological competitors in language mediated visual attention varies over the course of development (Mani & Huettig, submitted). This evidence has been argued to support the hypothesis that over the course of development the contribution of information from distinct modalities alters (Robinson and Sloutsky, 2004). By charting sensitivity to semantic and phonological competitors over the course of development in the MIM model, in which all modalities contribute equally throughout development, I test whether such additional architectural constraints are necessary in order to capture the developmental pattern observed.

Distinctions in the time course dynamics of fixations displayed towards distinct competitor types when placed in the same scene of visual world studies has been used to support theoretical descriptive accounts regarding modularity and casacading of information flow within language and visual processing systems (Altmann & Kamide, 2007; Altmann & Mirkovic, 2009; Huettig, Olivers & Hartsuiker, 2011; Huettig, Mishra & Olivers, 2012). I test

whether an interactive architecture such as that implemented in MIM is able to generate such features of language mediated visual attention, and if so by probing its internal processing I examine which properties of the system drive these observed effects.

Ambiguity in language is ubiquitous yet it is rarely harmful to effective communication (Piantadosi et al., 2012; Wasow & Arnold, 2003; Wasow et al., 2005; Roland, Elman & Ferreira, 2006; Ferreira, 2008; Jaeger, 2010). This suggests that the language processing system efficiently integrates extra-linguistic information available in the surrounding multimodal landscape with the speech signal it receives to resolve such ambiguity. Models of spoken word recognition however frequently overlook this multimodal aspect of speech processing (McClelland & Elman, 1986; Luce et al., 2000 Norris & McQueen, 2008; Scharenborg & Boves, 2010), therefore comparatively little is known about the structure of the system that supports this integration and its temporal characteristics. Within chapter 4 I investigate these issues by examining the ability of MIM to capture the interaction of visual information, semantic information and phonological information carried in the rhyme of words during spoken word processing.

Studies of language mediated visual attention have demonstrated that phonological rhyme competitors attract attention more than unrelated items (Allopenna et al., 1998; McQueen & Huettig, 2012; McQueen & Viebahn, 2007). However such effects have only been observed under heavily constrained laboratory conditions and in which phonology is the only dimension in which items in the display and the spoken target word are related. It is unknown whether phonological rhyme still exerts an influence on language mediated visual attention when other sources of information (e.g. visual or semantic) are also available to map between auditory and visual streams. Measuring sensitivity to such items provides an indirect measure of the influence of information carried in the rhyme of words during day to day spoken word processing when information from semantic or visual modalities may also be available to constrain spoken word processing. Further, serial, cascaded and parallel architectures which may support such multimodal integration make distinct predictions regarding the effect of this additional competition on phonological rhyme effects. Should processing proceed serially with the words full phonological form processed in full prior to influences from semantic or visual dimensions then rhyme effects should not be affected by the presence of visual or semantic competitors. However, in parallel and cascaded models visual and semantic information is available to affect processing before processing of information carried in the rhyme of the word is complete. Therefore, such models are likely to predict that

the presence of visual and semantic competitors will alter the nature of observed rhyme effects. This is tested in chapter 4 of this thesis. I first use MIM to generate predictions regarding the distribution of gaze toward phonological rhyme competitors in the presence and absence of visual and semantic competitors. This provides a prediction of behaviour given an architecture of spoken word recognition that facilitates the parallel integration of visual, semantic and phonological information. Predictions of the model are then tested in two visual world studies that expose participants to the conditions simulated in the model.

Having established in chapters 2 – 4 MIM's adequacy as a model of language mediated visual attention, in chapter 5 I move on to using it to test theories of the effects of literacy on language processing. Both computational and behavioural studies have demonstrated a connection between exposure to literacy training on alphabetic orthographies and the fidelity of phonological representations of words (Dijkstra, Roelofs & Fieuws, 1995; Chereau, Gaskell & Dumay, 2007; Hulme, Bowyer-Crane, Carroll, Duff, & Snowling, 2012; Kolinsky, Pattamadilok & Morais, 2012; Ventura, Morais, Pattmadilok & Kolinsky, 2004; Ziegler and Farrand, 1998), with data and theoretical models (Muneaux & Ziegler, 2004; Taft & Hambly, 1985; Taft, 2006; Ziegler & Goswami, 2005) suggesting phonological representations become more fine grained. However, such data is largely derived from explicit tasks in which participants explicitly manipulate the phonological structure of words. In contrast, comparatively little is known about the extent to which literacy training effects the structure of phonological processing during online speech processing (Reis & Castro-Caldas, 1997; Serniclaes et al., 2005; Huettig, Singh & Mishra, 2011). Literacy has also been linked to improved performance on a wide range of other cognitive tasks (Reis, Guerriero & Petersson, 2003; Kosmides, Tsapkini, Folia, Vlahou & Kiosseoglou, 2004; Reis & Castro-Caldes, 1997; Szwed, Ventura, Querido, Cohen and Dehaene, 2012; Bramao, Mendonca, Faisca, Ingvar, Petersson & Reis, 2007; Olivers, Huettig, Singh & Mishra, 2014) as has increased general processing speed (Kail & Salthouse, 1994; Li, Lindenberger, Hommel, Aschersleben, Prinz, & Baltes 2004; Salthouse, 2005). Therefore proposals have been made for a link between modulation of cognitive efficiency and literacy training (Stoodley & Stein, 2006; Pujol, et al., 2006). Recent studies of language mediated visual attention revealed distinctions in sensitivity to phonological competitors and semantic competitors between high literate and low literate populations (Huettig, Singh & Mishra, 2011). In chapter 5 I determine whether such effects were consistent with differences in cognitive efficiency and/or granularity of phonological representations recruited during online spoken word processing by

manipulating these parameters within MIM and observing their effects on the behaviour of the model.

In chapter 5 I simulate the effects of literacy by manipulating the structure of representations or processing characteristics within the model directly, I therefore do not capture the process by which exposure to orthographic mapping leads to changes to the language processing system. In chapter 6 I explore such emergent effects over the course of reading development by manipulating within a computational model of reading (compatible with the hub and spoke framework implemented in chapters 2 – 5: see Dilkina et al., 2008; 2010) the structure of the orthography on which networks are trained.

There is dramatic variation across the worlds orthographic systems in the way they encode a languages phonological and semantic structure (i.e. their semantic or phonological transparency). For all of these systems over the course of literacy training readers learn to integrate this orthographic structure with the existing language processing system. Studies of the effects of orthographic transparency suggest that the structure of the orthographic system is likely to have major implications for how the reading system operates and is integrated with existing language processing networks. For example behavioural and neural imaging data suggests orthographic transparency impacts on reading acquisition (Snowling & Hulme, 2005; Goswami, Gombert, & De Barrera, 1998; Bruck, Genesee and Caravolas, 1997; Nag, 2007; Asfaha et al., 2009), the division of labour across components of the reading system (Paulesu et al., 2000; Kiyosawa et al., 1995) and the structure of phonological representations (Perre et al., 2009; Pattamadilok et al., 2010). However, such studies are compromised due to co-varying linguistic, socio-economic and socio-cultural factors. Within chapter 6 of this thesis I train emergent neural network implementations of the triangle model of reading (Seidenberg & McClelland, 1989; Harm & Seidenberg, 2004) on a range of orthographic systems (Comrie, 2013: alphabetic, consonantal, syllabic, alphasyllabic, logographic) that represent the range of the world's writing systems while holding phonological and semantic structure constant. By manipulating orthographic transparency in this way I'm able to isolate its effects and avoid confounds that have plagued previous behavioural and neural imaging studies.

The literature is divided on the properties that define a universal model of reading (Frost, 2012; Perfetti et al., 2005). Using these models I test the adequacy of the triangle model as such a model by examining its ability to support reading across this comprehensive range of

orthographies. Models are also assessed on their phonological decoding and reading comprehension acquisition rates. This allows us to test the hypothesis that transparency aids phonological decoding acquisition (Goswami, Gombert & De Barrera, 1998; Seymour, Aro & Erskine, 2003; Bruck, Genesee & Caravolas, 1997; Hanley, Masterson, Spencer & Evans, 2004) and make predictions of effects of transparency on comprehension which remains an underexplored issue (Seidenberg, 2013). Neural imaging and computational modelling data suggests that transparency also affects the flow of activation across neural pathways within the reading system (Paulesu et al., 2000; Kiyosawa et al., 1995). By recording the flow of activation across direct and indirect paths connecting phonological, semantic and orthographic processing components within implemented models I test the extent to which semantic and phonological transparency is able to modulate the distribution of activation across such paths. Further, as described earlier in this introduction both behavioural (see Zeigler & Goswami, 2005; Morais & Kolinsky, 2001; Petersson, Ingvar & Reis, 2009) and neural data (Dehaene et al., 2010; Perre et al., 2009; Pattamadilok et al., 2010) has been put forward in support of theoretical arguments that orthographic transparency affects the structure of phonological representations. I examine whether processing within the implemented interactive models of reading detailed in chapter 6 are consistent with this body of data. This is done by testing whether differences in correspondence between orthography and phonology, and orthography and semantics affect the structure of phonological processing. I also test whether models' predict novel, analogous effects in the semantic domain by also testing whether phonological and semantic transparency affects the structure of semantic processing.

Finally, in chapter 7 I summarise the results from each chapter and discuss the conclusions that can be drawn from this collective group of studies.

Reading Guide

Chapters 2 – 6 of this thesis have been written in order for them to stand alone as independently publishable papers, therefore there is some repetition of both the description of the literature and method sections.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4), 419-439.

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, 57(4), 502-518.

Altmann, G., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33(4), 583-609.

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, 166(3-4), 559-571.

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica*, 137(2), 181-189.

Asfaha, Y. M., Kurvers, J., & Kroon, S. (2009). Grain size in script and teaching: Literacy acquisition in Ge'ez and Latin. *Applied Psycholinguistics*, 30(04), 709-724.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, 13, 99-102.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, 59, 617-645.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences*, 22(04), 637-660.

Berkeley, G. (1709). *An essay towards a new theory of vision*. Aaron Rhames.

Bramao, I., Mendonca, A., Faisca, L., Ingvar, M., Petersson, K. M., & Reis, A. (2007). The impact of reading and writing skills on a visuo-motor integration task: A comparison between illiterate and literate subjects. *Journal of the International Neuropsychological Society*, 13(2), 359-364.

Brennan, C., Cao, F., Pedroarena-Leal, N., McNorgan, C., & Booth, J. R. (2013). Reading acquisition reorganizes the phonological awareness network only in alphabetic writing systems. *Human brain mapping*, 34(12), 3354-3368.

Brown-Schmidt, S., & Konopka, A. E. (2008). Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition*, 109(2), 274-280.

Bruck, M., Genesee, F., & Caravolas, M. (1997). A cross-linguistic study of early literacy acquisition. *Foundations of reading acquisition and dyslexia: Implications for early intervention*, 145-162.

Cao, F., Khalid, K., Lee, R., Brennan, C., Yang, Y., Li, K., Bolger, D. J. & Booth, J. R. (2011). Development of brain networks involved in spoken word processing of Mandarin Chinese. *NeuroImage*, 57(3), 750-759.

Caramazza, A. (1997). How many levels of processing are there in lexical access?. *Cognitive neuropsychology*, 14(1), 177-208.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, 10(7), 335-344.

Chéreau, C., Gaskell, M. G., & Dumay, N. (2007). Reading spoken words: Orthographic effects in auditory priming. *Cognition*, 102(3), 341-360.

Cheung, H., Chen, H. C., Lai, C. Y., Wong, O. C., & Hills, M. (2001). The development of phonological awareness: Effects of spoken language experience and orthography. *Cognition*, 81(3), 227-241.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1), 204.

Comrie, B. (2013). Writing Systems. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/141, Accessed on 2014-05-07.)

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6(1), 84-107.

Cutler, A. and Norris, D. G. (1979) Monitoring sentence comprehension. In W. E. Cooper and E. C. T.Walker (eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. Hillsdale, N J, Erlbaum.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic bulletin & review*, 12(3), 453-459.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J. & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science*, 330(6009), 1359-1364.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3), 283.

Dijkstra, T., Roelofs, A., & Fieuws, S. (1995). Orthographic effects on phoneme monitoring. *Canadian Journal of Experimental Psychology*, 49(2), 264.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, 25(2), 136-164.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010). Are there mental lexicons? The role of semantics in lexical decision. *Brain research*, 1365, 66-81.

Eggert, G. H. (1977). Wernicke's works on aphasia: a sourcebook and review. Mouton de Gruyter.

Elman, J. L. (Ed.). (1996). *Rethinking innateness: A connectionist perspective on development* (Vol. 10). MIT press.

Elman, J. L. (2005). Connectionist models of cognitive development: where next?. *Trends in cognitive sciences*, 9(3), 111-117.

Ferreira, V. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation*, 49, 209–246.

Ferreira, F., & Tanenhaus, M. K. (2007). Introduction to the special issue on language–vision interactions. *Journal of Memory and Language*, 57(4), 455-459.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT press.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6(2), 78–84.

Frost, R. (2012). A universal approach to modeling visual word recognition and reading: Not only possible, but also inevitable. *Behavioral and Brain Sciences*, 35(05), 310-329.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. J*ournal of Experimental Psychology: Human Perception and Performance*, 13(1), 104.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12(5-6), 613-656.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory?. *Trends in cognitive sciences*, 10(6), 278-285.

Gibbs, R. W. (2006). Metaphor interpretation as embodied simulation. *Mind & Language*, 21(3), 434-458.

Goswami, U., Gombert, J. E., & de Barrera, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French, and Spanish. *Applied Psycholinguistics*, 19(01), 19-52.

Hanley, R., Masterson, J., Spencer, L., & Evans, D. (2004). How long do the advantages of learning to read a transparent orthography last? An investigation of the reading skills and reading impairment of Welsh children at 10 years of age. *The Quarterly Journal of Experimental Psychology: Section A*, 57(8), 1393-1410.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, 106(3), 491-528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review*, 111(3), 662-720.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1), 335-346.

Ho, C. S. H., & Bryant, P. (1997). Phonological skills are important in learning to read Chinese. *Developmental Psychology*, 33(6), 946.

Hollich, G.J., Hirsh-Pasek, K., Golinkoff, R.M., Brand, R.J., Brown, E., Chung, H.L., Hennon, E., Rocroi, C., & Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 65, 1-135.

Huang, H. S., & Hanley, J. R. (1995). Phonological awareness and visual skills in learning to read Chinese and English. *Cognition*, 54(1), 73-98.

Huang, H. S., & Hanley, J. R. (1997). A longitudinal study of phonological awareness, visual skills, and Chinese reading acquisition among first-graders in Taiwan. *International Journal of Behavioral Development*, 20(2), 249-268.

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: evidence from real-time spoken language comprehension. *Developmental psychology*, 45(6), 1723.

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm.*Cognition*, 96(1), B23-B32.

Huettig, F., & Altmann, G. T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, 15(8), 985-1018.

Huettig, F., & Janse, E. (in press). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. L*anguage, Cognition and Neuroscience*.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, 57(4), 460-482.

Huettig, F., Mishra, R. K., & Olivers, C. N. (2012). Mechanisms and representations of language-mediated visual attention. *Frontiers in psychology*, 2.394.

Huettig, F., Olivers, C. N., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta psychologica*, 137(2), 138-150.

Huettig, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. T. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta psychologica*, 121(1), 65-80.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137(2), 151-171.

Huettig, F., Singh, N., & Mishra, R. K. (2011). Language-mediated visual orienting behavior in low and high literates. *Frontiers in psychology*, 2.

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The Causal Role of Phoneme Awareness and Letter-Sound Knowledge in Learning to Read Combining Intervention Studies With Mediation Analyses. *Psychological Science*, 23(6), 572-577.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1), 23-62.

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta psychologica*, 86(2), 199-225.

Katz, L., & Feldman, L. B. (1981). Linguistic coding in word recognition: Comparisons between a deep and a shallow orthography. In A. M. Lesgold & C. A. Perfetti (Eds.), *Interactive Processes in Reading*. Hillsdale, NJ: Erbaum, 85-106.

Kello, C. T., & Plaut, D. C. (2000). Strategic control in word reading: evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(3), 719-750.

Kirkham, N.Z. (2010) Altogether now: Learning through multiple sources. In S.P. Johnson (Ed.), *Neoconstructivism: The new science of cognitive development*. New York: Oxford University Press.

Kiyosawa, M., Itoh, M., Nakagawa, Y., Kobayashi, N., Tamai, M., 1995. Effect of kanji and kana reading on cerebral blood flow patterns measured by PET. *Jpn. J. Ophthalmol*. 39, 198–205.

Kolinsky, R., Pattamadilok, C., & Morais, J. (2012). The impact of orthographic knowledge on speech processing. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, (63), 161-186.

Kosmidis, M. H., Tsapkini, K., Folia, V.., Vlahou, C. H., & Kiosseoglou, G.. (2004). Semantic and phonological processing in illiteracy. *Journal of the International Neuropsychological Society*, 10(6), 818-827.

Kuhl, P. K. (2000). Language, mind, and brain: Experience alters perception. *The new cognitive neurosciences*, 2, 99-115.

Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and differentiation in semantic memory. *Annals of the New York Academy of Sciences*, 1124(1), 61-76.

Levelt, W. J. (1999). Producing spoken language: A blueprint of the speaker. In T*he neurocognition of language* (pp. 83-122). Oxford University Press.

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in cognitive sciences*, 13(11), 470-478.

Li, S. C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science*, 15(3), 155-163.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics*, 62(3), 615-625.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review,* 101(4), 676.

Majid, A., Bowerman, M., Kita, S., Haun, D. B., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in cognitive sciences*, 8(3), 108-114.

Mani, N., & Huettig, F. (submitted). The changing dynamics of word-referent mapping across development.

Mani, N., Johnson, E., McQueen, J. M., & Huettig, F. (2013). How yellow is your banana? Toddlers' language-mediated visual search in referent-present tasks. *Developmental psychology*, 49(6), 1036.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8(1), 1-71.

McBride-Chang, C., Bialystok, E., Chong, K. K., & Li, Y. (2004). Levels of phonological awareness in three cultures. *Journal of Experimental Child Psychology*, 89(2), 93-111.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, 18(1), 1-86.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14(8), 348-356.

McClelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive science*, 38(6), 1139-1189.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). Parallel distributed processing. Explorations in the microstructure of cognition, 2, 216-271.

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive psychology*, 60(1), 1-39.

McNorgan, C., Reid, J., & McRae, K. (2011). Integrating conceptual knowledge within and across representational modalities. *Cognition*, 118(2), 211-233.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognizeda). *The Journal of the Acoustical Society of America*, 131(1), 509-517.

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *The Quarterly Journal of Experimental Psychology*, 60(5), 661-671.

McQueen, J. M., Dahan, D., & Cutler, A. (2003). Continuity and gradedness in speech processing. Phonetics and phonology in language comprehension and production: Differences and similarities, 39-78.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & cognition*, 37(7), 1026-1039.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of memory and language*, 59(4), 475-494.

Moore, C., Angelopoulos, M., & Bennett, P. (1999). Word learning in the context of referential and salience cues. *Developmental Psychology*, 35(1), 60-68.

Morais, J., & Kolinsky, R. (2001). The literate mind and the universal human mind. In Dupoux, E., & Mehler, J. (Eds). Lang*uage, brain and cognitive development: Essays in Honor of Jacques Mehle*. MIT, Cambridge, Mass, 463-480.

Morton, J. (1969). Interaction of information in word recognition. Psychological review, 76(2), 165.

Munakata, Y., & McClelland, J. L. (2003). *Connectionist models of development. Developmental Science*, 6(4), 413-429.

Muneaux, M., & Ziegler, J. (2004). Locus of orthographic effects in spoken word recognition: Novel insights from the neighbour generation task. *Language and Cognitive Processes*, 19(5), 641-660.

Nag, S. (2007). Early reading in Kannada: The pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading*, 30(1), 7-22.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review*, 115(2), 357.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23(03), 299-325.

Olivers, C. N. L., Huettig, F., Singh, J. P., & Mishra, R. K. (2014). The influence of literacy on visual search. *Visual Cognition*, 22(1), 74-101.

Pattamadilok, C., Knierim, I. N., Duncan, K. J. K., & Devlin, J. T. (2010). How does learning to read affect speech perception? *The Journal of Neuroscience*, 30(25), 8435-8444.

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S. F., Cotelli, M. Cossu, G., Corte, F., Lorusso, M., Pesenti, S., Gallagher, A., Perani, D., Price, C., Frith, C. D., & Frith, U. (2000). A cultural effect on brain function. *Nature neuroscience*, 3(1), 91-96.

Perfetti, C. A., Liu, Y., & Tan, L. H. (2005). The lexical constituency model: some implications of research on Chinese for general theories of reading. *Psychological Review*, 112(1), 43-59.

Perre, L., Pattamadilok, C., Montant, M., & Ziegler, J. C. (2009). Orthographic effects in spoken language: On-line activation or phonological restructuring?. *Brain research*, 1275, 73-80.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, 114, 273-315.

Perry, C., Ziegler, J. C., & Zorzi (2010). Beyond single syllables: Large-scale modelling of reading aloud with the connectionist dual process (CDP++) model. *Cognitive Psychology*, 61, 2, 106-151.

Petersson, K. M., Ingvar, M., & Reis, A. (2009). Language and literacy from a cognitive neuroscience perspective. In D. Olsen, & N. Torrance (Eds.), *Cambridge handbook of literacy* (pp. 152-181). Cambridge: Cambridge University Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3), 280-291.

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology*, 19(7), 603-639.

Prinz, J. (2002). *Furnishing the mind: concepts and their perceptual basis*. Cambridge, MA: MIT Press.

Pujol, J., Soriano-Mas, C., Ortiz, H., Sebastian-Galles, N., Losilla, J. M., & Deus, J. (2006). Myelination of language-related areas in the developing brain. *Neurology*, 66(3), 339-343.

Pulvermüller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. *Brain and language*, 110(2), 81-94.

Read, C., Yun-Fei, Z., Hong-Yin, N., & Bao-Qing, D. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition*, 24(1), 31-44.

Reis, A., & Castro-Caldas, A. (1997). Illiteracy: A cause for biased cognitive development. *Journal of the International Neuropsychological Society*, 3(05), 444-450.

Reis, A., Guerreiro, M. & Petersson, K. M. (2003). A sociodemographic and neuropsychological characterization of an illiterate population. *Applied Neuropsychology*, 10(4), 191–204.

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development, 75*(5), 1387-1401.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science, 38*(6), 1024-1077.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological Review, 111*(1), 205-235.

Roland, D., Elman, J. L., & Ferreira, V. S. (2006). Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*, 98(3), 245-272.

Rumelhart, D. E., & McClelland, J. L., & the PDP Research Group (1986). Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: Foundations & volume II: Psychological and biological models. Cambridge, MA: MIT Press.

Salthouse, T. A. (2005). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology*, 19(4), 532-545.

Scharenborg, O., & Boves, L. (2010). Computational modelling of spoken-word recognition processes: Design choices and evaluation. *Pragmatics & Cognition*, 18(1), 136-164.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4), 523.

Seidenberg, M. S. (2013). The science of reading and its educational implications. *Language Learning and Development*, 9, 331-360.

Serniclaes, W., Ventura, P., Morais, J., & Kolinsky, R. (2005). Categorical perception of speech sounds in illiterate adults. *Cognition*, 98, B35–B44.

Seymour, P. H., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94, 143-174.

Shallice, T. (1988). Specialisation within the semantic system. *Cognitive neuropsychology*, 5(1), 133-142., 1988;

Shu, H., Peng, H., & McBride-Chang, C. (2008). Phonological awareness in young Chinese children. *Developmental Science*, 11(1), 171-181.

Sirois, S., Spratling, M., Thomas, MSC., Westermann, G., Mareschal, D., & Johnson, M. (2008). Précis of Neuroconstructivism. *Behavioral and Brain Sciences*, 31, 321-331.

Snowling, M. J., & Hulme, C. (Eds.). (2005). *The science of reading: A handbook*. Oxford, UK: Blackwell.

Stoodley, C. J., & Stein, J. F. (2006). A processing speed deficit in dyslexic adults? Evidence from a peg-moving task. *Neuroscience letters*, 399(3), 264-267.

Szwed, M., Ventura, P., Querido, L., Cohen, L., & Dehaene, S. (2012). Reading acquisition enhances an early visual process of contour integration. *Developmental science,* 15(1), 139-149.

Taft, M. (2006). Orthographically influenced abstract phonological representation: Evidence from non-rhotic speakers. *Journal of psycholinguistic research,* 35(1), 67-78.

Taft, M., & Hambly, G. (1985). The influence of orthography on phonological representations in the lexicon. *Journal of Memory and Language*, 24(3), 320-335.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632-1634.

Ueno, T., & Lambon Ralph, M. A. (2013). The roles of the "ventral" semantic and "dorsal" pathways in conduite d'approche: A neuroanatomically-constrained computational modeling investigation. *Frontiers in Human Neuroscience*, 7, 422.

UNESCO Institute for Statistics. (2013). *Adult and Youth Literacy Fact Sheet*, Montreal: UNESCO.

Van de Velde, M., Meyer, A. S., & Konopka, A. E. (2014). Message formulation and structural assembly: describing "easy" and "hard" events with preferred and dispreferred syntactic structures. *Journal of Memory and Language*, 71(1), 124-144.

Ventura, P., Morais, J., Pattamadilok, C., & Kolinsky, R. (2004). The locus of the orthographic consistency effect in auditory word recognition. *Language and Cognitive Processes*, 19(1), 57-95.

Wasow, T., & Arnold, J. (2003). Post-verbal constituent ordering in English. *Determinants of Grammatical Variation in English*, 119–154.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*. Stanford: CSLI Publications.

Westermann, G., & Ruh, N. (2012). A neuroconstructivist model of past tense development and processing. *Psychological Review*, 119, 649-667.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4), 625-636.

Wright, B. A., Lombardino, L. J., King, W. M., Puranik, C. S., Leonard, C. M., & Merzenich, M. M. (1997). Deficits in auditory temporal and spectral resolution in language-impaired children. *Nature*, 387, 176-178.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 1-14.

Yee, E., Overton, E., & Thompson-Schill, S. L. (2009). Looking for meaning: Eye movements are sensitive to overlapping semantic features, not association. *Psychonomic Bulletin & Review, 16*(5), 869-874.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70*(13), 2149-2165.

Yurovsky, D., Boyer, T. W., Smith, L. B., & Yu, C. (2013). Probabilistic cue combination: less is more. *Developmental Science*, *16*(2), 149-158.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin*, 131(1), 3.

Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review*, 5(4), 683-689.

Chapter 2

A multimodal integration model (MIM) of language-mediated visual attention[1]

Abstract

Language-mediated visual attention describes the interaction of two fundamental components of the human cognitive system, language and vision. Within this paper we present a shared resource model of language-mediated visual attention that offers a description of the information and processes involved in this complex multimodal behaviour and a potential explanation for how this ability is acquired. We demonstrate that the model is not only sufficient to account for the experimental effects of Visual World Paradigm studies but also that these effects are emergent properties of the architecture of the model itself, rather than requiring separate information processing channels or modular processing systems. The model provides an explicit description of the connection between the modality-specific input from language and vision and the distribution of eye gaze in language mediated visual attention. The paper concludes by discussing future applications for the model, specifically its potential for investigating the factors driving observed individual differences in language mediated eye gaze.

[1] *Adapted from Smith, A. C., Monaghan, P., & Huettig, F. (2013). An amodal shared resource model of language-mediated visual attention. Frontiers in Psychology. 4:528.*

1. Integrative processing in a model of language mediated visual attention

Language mediated visual attention

Within daily communicative interactions a vast array of information sources have to be integrated in order to understand language and relate it to the world around the interlocutors. Such multimodal interactions within the speaker and listener have been shown to be vital for language development (Bloom, 2000; Mani, Johnson, McQueen, & Huettig, 2013; Markman, 1994; Monaghan & Mattock, 2012) as well as for adult sentence and discourse processing (Anderson, Chiu, Huette, & Spivey, 2011; Huettig, Olivers & Hartsuiker, 2011; Lupyan, 2012). Eye gaze has been used to demonstrate the nature of the processes supporting online integration of linguistic and visual information (Halberda, 2006; Huettig, Mishra, & Olivers, 2012). Such observations of eye gaze have opened up the possibility to investigate how multiple sources of information, within the environment and within the language signal, interact in the human cognitive system. We begin by describing the observed properties of eye gaze behaviour that have informed our understanding of the representations and processes involved in language – vision interactions. We then present a computational model of language mediated visual attention that implements the representations and processes identified within a parsimonious neural network architecture. Finally, we demonstrate that many of the characteristic features of language mediated eye gaze can be captured by the emergent properties of this parsimonious architecture and therefore do not necessitate separate information processing channels or modular processing systems.

One influential paradigm for measuring language and vision interactions is the Visual World Paradigm (VWP; Copper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), in which participants are presented with a visual display comprising a set of objects and/or actors whilst hearing an auditory stimulus and during this period their eye gaze is recorded. Although eye gaze is a measure of overt attention and thus not a direct reflection of linguistic processing, the VWP has been utilised largely to investigate questions that explore how the cognitive system processes spoken language (see Huettig, Rommers & Meyer, 2011, for review). A few studies however have investigated multimodal interactions. Such studies tend to focus on how eye gaze alters as the auditory stimulus unfolds and how varying the relationships between objects in the display can highlight which modalities of information are implicated at varying time points in language processing.

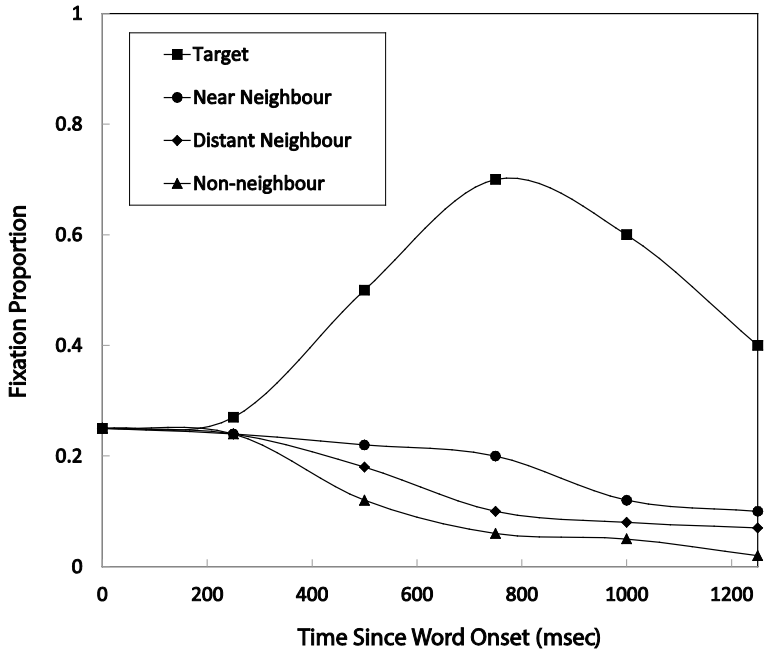

Figure 1: Figure adapted from Mirman and Magnuson (2009). Figure displays approximate fixation proportions for targets, near semantic neighbours, distant semantic neighbours and unrelated items displayed by participants in Mirman & Magnuson (2009).

Many visual world studies have demonstrated that eye gaze can be modulated by phonological relationships between items presented in the visual display and spoken target words. Allopenna, Magnuson, and Tanenhaus (1998), for instance, showed that when hearing a target word (e.g., "beaker") participants looked more towards items in the display whose names overlapped phonologically with the target word either in initial (e.g., beetle) or final (e.g., speaker) positions, than items that were not related phonologically (e.g., carriage) to the spoken target word. They found that, relative to unrelated items, there was increased fixation of phonological competitors. Furthermore, fixations to onset competitors occurred earlier than those to rhyme competitors and the probability of fixating onset competitors was greater than the probability of fixating rhyme competitors.

Visual relationships between items have also been shown to influence fixation behaviour (Dahan & Tanenhaus, 2005; Huettig & Altmann, 2007). Dahan and Tanenhaus (2005) presented scenes containing a target (e.g., a snake), a visual competitor (e.g., a rope) and two unrelated distractors (e.g., a couch and an umbrella), while Huettig and Altmann (2007) presented scenes without a visual depiction of the target but with a visual competitor and three unrelated distractors. Thus, items within the display that shared visual features

associated with the spoken target word, yet whose names did not overlap phonologically with the target word, attracted greater fixation than unrelated items.

Another dimension in which relationships between visually displayed items and spoken target words has been shown to modulate eye gaze is that of semantics. Huettig and Altmann (2005) and Yee and Sedivy (2006) demonstrated that items that share semantic (but not visual or phonological) relationships with target words are fixated more than unrelated items. Yee and Sedivy (2006) presented displays containing a target item (e.g., lock), a semantically related item (e.g., key) and two unrelated distractors. Similarly, Huettig and Altmann (2005) presented scenes containing both a target (e.g., piano) and a semantic competitor (e.g., trumpet) or scenes containing only a semantic competitor (e.g., only the trumpet) and unrelated items. In both target present and target absent conditions increased fixations of semantically related items were observed. Post-hoc analyses revealed that the likelihood of fixation was proportional to the degree of semantic overlap as measured by feature production norms (cf. Cree and McRae, 2003) and corpus-based measures of word semantics (Huettig, Quinlan, McDonald, & Altmann, 2006). Further evidence for a relationship between semantic overlap and eye gaze is provided by Mirman and Magnuson (2009) who directly tested the gradedness of semantic overlap. They presented scenes containing a target item (e.g., bus) paired with either a near semantic neighbour (e.g., van) or a distant semantic neighbour (e.g., bike) and two unrelated items (e.g., ball). The likelihood of fixating each item was predicted by the level of semantic overlap, with near semantic neighbours fixated with greater probability than far semantic neighbours, while both were fixated with lower probability than targets and greater probability than distractors (see Figure 1).

In order to probe the relationships between previously observed phonological, visual and semantic word level effects in the VWP, Huettig and McQueen (2007) presented scenes containing phonological onset, semantic and visual competitors in addition to an unrelated distractor. They observed distinct patterns of fixation for each competitor type, with participants initially looking more towards phonological onset competitors before later displaying greater fixation of visual and semantic competitors. From these results they concluded that language mediated visual attention is determined by matches between information extracted from the visual display and speech signal at phonological, visual and semantic levels of processing.

Taken together, this significant body of evidence shows that visual, semantic and phonological information is co-activated and integrated during spoken word processing. However, the nature of the information and mechanisms involved in visual world and language processing interactions are as yet underspecified (Huettig, Olivers et al., 2011; Huettig, Mishra & Olivers, 2012). How is information activated within one modality integrated with information activated within another, what form does this information take, how does such information interact and how is this interaction connected to eye gaze behaviour? There are two principle possibilities for interactions to occur: They may be a consequence of modality specific systems interacting via direct connections; alternatively, interactions may occur as a consequence of amodal shared resources facilitating interaction between the various information modalities (Lambon Ralph & Patterson, 2008; Plaut, 2002). Computational implementation of theoretical models offers a means of testing their plausibility and often provides a means of probing aspects of theoretical models that may lie beyond the reach of behavioural studies. The VWP provides a high degree of experimental control that offers a well constrained environment in which models can operate. Models of the processes involved in performing VWP tasks force researchers to define explicitly how information carried in the visual and auditory stimuli is connected to distributions of eye gaze.

In this paper, we first present previous modelling approaches that have accounted for the various VWP results presented above before elaborating the modular versus shared-resource computational approaches to multimodal information processing. We then present our model of the shared resource account of language mediated visual attention and demonstrate that it is not only sufficient to account for the experimental effects of VWP studies but also that these effects are emergent properties of the architecture of the model itself.

Previous models of language mediated visual attention

Most previous models of the VWP have focused on explaining interactions between vision and a single feature of language processing. For instance, Allopenna et al., (1998) chose TRACE (McClelland & Elman, 1986) to simulate the mechanisms driving differences in the effect of phonological onset and rhyme overlap. TRACE is a continuous mapping model of speech perception, implemented in an interactive activation network that hierarchically processes speech at the level of phonemic features, phonemes and words. The model successfully replicated the contrasting patterns of fixation displayed by participants toward

onset and rhyme competitors and offered explanation for contrasts between the location of overlap and its influence on eye gaze. However, the model focuses purely on phonological processing and therefore as a model of language mediated visual attention it provides no description of the role other information sources play in this process.

Magnuson, Tanenhaus, Aslin and Dahan (2003) further examined the mechanisms underlying observed cohort and rhyme effects. They demonstrated that differences in sensitivity to both cohort and rhyme competitors displayed by adults over the course of word learning could be captured in the emergent behaviour displayed by an SRN (Elman, 1990) trained to map between phonetic features and localist word level representations. Unlike TRACE, in which connection weights were fixed by the modeller, connection weights within the SRN were adjusted using an error based learning algorithm. This not only reduces the number of parameters directly manipulated by the modeller and therefore the number of assumptions underpinning the model but also allowed authors to chart model behaviour over the course of word learning. Using this approach they were able to demonstrate that a fundamental difference between adult and child lexical representations was not required to explain differences in sensitivity to rhyme and cohort competitors. Instead such differences were captured by their model due to the strengthening of lexical representations over the course of word learning. Again however, the focus of this work is on aspects of phonological processing in the VWP. Therefore, as a model of language mediated visual attention it ignores the role of other knowledge types in this process.

Similarly, Mirman and Magnuson (2009) used the attractor network of Cree et al., (1999) to simulate the graded effect of semantic competitors influencing eye gaze. The network consisted of a word form input layer and semantic feature output layer. The model was trained to map 541 words onto their corresponding semantic features derived from feature norming studies. However, as in the case of Magnuson et al. (2003) and TRACE, such models offer representation of items from only a single information source (phonological or semantic similarity) and therefore are unable to account for the full range of intermodal effects demonstrated in the VWP. Also, none of these models offer a description of how information activated by distinct visual and auditory sources can be combined to influence fixation behaviour. They therefore do not provide a comprehensive model of the word level effects observed in the VWP.

There have, however, been some notable models of multimodal processing in VWP (Spivey, 2008; Mayberry, Crocker & Knoeferle, 2009; Kukona & Tabor, 2011). Spivey (2008) extended TRACE to incorporate visual processing, by connecting lexical activations in TRACE to a normalised recurrent localist attractor network that represented the presence or absence of items within the visual environment. However, in using localist visual representations the model lacks depth of representation in the visual modality to capture subtle relationships between items known to influence fixation behaviour in VWP, such as visual similarity effects.

Mayberry et al. (2009) also provided a model of the interaction between visual and linguistic information in the VWP. Their connectionist model (CIANET) displays emergent properties that capture sentence level effects such as case role interpretation. A potential weakness of the model is its use of the same form of representation to encode both visual and linguistic information, thereby masking potential distinct effects of visual versus linguistic similarities. A further weakness of both CIANET and Spivey (2008) is that neither provide representation at the word level in a semantic dimension, although we know from previous VWP studies that items can differ in both visual and phonological dimensions yet still share semantic properties that influence eye gaze behaviour.

Finally, Kukona and Tabor (2011) presents a dynamical systems model of eye gaze in VWP in which localist representations at phonological, lexical-semantic, cross-word and action-space layers interact in a hierarchically structured network. Visual information is modelled by the presence or absence of its corresponding representations within the network. By representing items at this level of abstraction their model is unable to capture complex relationships between representations in the same modality. It seems then that none of the current multimodal models that have been used to explicitly model VWP data offer sufficient depth of representation in the multiple modalities involved to capture the subtle relationships between items shown to influence eye gaze at the word level in VWP.

Yet, previous models and their success in replicating individual VWP data sets have provided valuable insight into the type of architecture capable of supporting language-mediated visual attention. The architecture must allow for competition at multiple levels of representation (Allopenna et al., 1998), allow both excitatory and inhibitory connections (Mirman & Magnuson, 2009), facilitate parallel activation of representations (Kukona & Tabor, 2011) and integrate information from multiple sources (Mayberry et al., 2009). Such integration

could be accomplished by connectivity between individual representational modalities, or via processing interconnectivity through a shared resource.

Modular versus shared-resource models

A framework able to capture the architectural features of language mediated visual attention identified in the previous section is the Hub-and-spoke (H&S) framework. H&S models are defined by a central resource (hub) that integrates and translates information between multiple modality specific sources (spokes). The framework arose as one side of a debate regarding the neural structures that support human conceptual and semantic knowledge. Lambon Ralph and Patterson (2008) compared two alternative theoretical models to account for visual and linguistic semantic processing in unimpaired and patient populations. One consisted purely of modality specific processing regions connected via direct connections, the second instead connected regions via a modality invariant central hub, the H&S model. The authors argue that although a web of direct connections may provide a simpler architectural solution, only a model that contains a central connecting hub offers a system capable of performing the multilevel nonlinear computations required for semantic generalisation and inference based on conceptual structure rather than surface similarities. There is also converging empirical evidence for both the existence of a semantic hub and its implementation in specific neural populations in the anterior temporal lobe (ATL). This evidence includes neuropsychological studies of patients suffering from semantic dementia (SD) (Lambon Ralph et al., 2010) who possess lesions in the ATL and display deficits in performance on tasks requiring semantic generalisation. Similarly, non-patient groups that experience artificial lesions in the ATL using rTMS (Pobric, Jefferies & Lambon Ralph, 2007) have reported similar deficits in performance on such tasks. Finally, neuroimaging studies (Vandenberghe et al., 1996), have observed activity in the ATL on tasks that require semantic generalisation. These data support the notion that a central resource that integrates modality specific information is a crucial component of the architecture supporting semantic processing.

Models that postulate integrative processing from multiple sources are embedded in a broader literature that has debated the inherence of sensory and motor systems to conceptual representations. Studies of “embodied cognition”, for instance, have made the case for the importance of motor and sensory systems for cognitive processing (e.g., Barsalou, Simmons, Barbey, & Wilson, 2003, but see Mahon & Caramazza, 2008). An important debate concerns

the format of mental representations with some proponents of the embodied cognition hypothesis suggesting that conceptual knowledge consists entirely of "representational codes that are specific to our perceptual systems" (Prinz, 2002, p.119). This contrasts with representational theories which assume that sensory and motor knowledge is amodal and abstracted away from modality-specific systems (e.g., Kintsch, 2008). A third view posits the existence of both amodal and modal representations providing an explanation of how we are able to acquire knowledge which goes beyond sensory and motor experience (Dove, 2009; Goldstone & Barsalou, 1998). This view is supported by recent demonstrations that co-activation of multimodal systems can be effectively simulated by models with an amodal shared resource (Monaghan & Nazir, 2009; Yoon, Heinke, & Humphreys, 2002). Given that activation in a spoke of a H&S model represents modality specific processing of an item, and activation within the hub may capture an items amodal properties, then the interaction of modal (spoke) and amodal (hub) representations is a potential consequence of the architecture of H&S models. A recent review of the mechanisms and representations involved in language-mediated visual attention (Huettig, Mishra & Olivers, 2012) concluded that the most promising theoretical model to date postulates that language mediated visual attention is dependent on a system in which both linguistic, non-linguistic and attentional information are all instantiated within the same coding substrate, which is required in order for information to be bound across modalities. The H&S framework offers a parsimonious solution by connecting modalities through a central processing hub.

Research examining the plausibility of alternative theoretical models of multimodal cognition has profited from testing their predictions using explicit neural network implementations of the H&S framework. In the following sections we detail the nature of these studies and how they have contributed to our understanding of the mechanisms that support semantic processing. We also identify the features of the Hub and Spoke framework that make it a valuable tool for modelling various aspects of multimodal cognition. We then test the framework's scope by using it as a foundation for a model of language mediated visual attention.

Insights from Hub & Spoke models

The H&S framework offers a parsimonious architecture in which single modality models can be drawn together to examine the consequences of multimodal interaction. Producing an explicit model of the mechanisms thought to underlie a given process allows one to test

theoretical positions and probe deeper the mechanisms that may be involved in a controlled and tractable manner.

The framework provides a single system architecture with only minimal initial assumptions on connectivity. As the systems architecture imposes minimal constraints on the flow of information within the network, emergent behaviour is largely driven by 1) representational structure and/or 2) the tasks or mappings performed by the system during the learning process. Therefore, within the framework the scope of such factors in driving emergent properties of complex multimodal systems can be examined largely independent of modality specific architectural constraints.

Two alternative means of exploring the role of representational structure are presented in previous H&S models. Plaut (2002) simulates impairments displayed by optic aphasics in an H&S model that mapped between distinct vision, action (gesturing), touch and phonological (naming) layers. The author takes a fundamentalist approach (see Kello & Plaut, 2000) ensuring he has total control over any relationships embedded in representations within or across modalities. This allows the study to isolate emergent properties driven by individual aspects of representational structure. In Plaut (2002) the variable manipulated was systematicity in representation between modalities. He embedded systematic mappings between tactile, vision and action representations while those between phonology and other modalities were arbitrary. This feature of representations allowed the model to capture key features of patient behaviour with the lack of systematicity in phonological representations leading to poor performance on naming tasks post lesioning.

In contrast, Rogers et al., (2004) (approach replicated in Dilkina, McClelland & Plaut 2008; 2010) employs a realist approach with representations derived from feature norming studies. Within the study deficits in semantic processing displayed by SD patients are modelled using an H&S framework. The model consisted of a visual layer connected via a central resource to a verbal descriptor layer comprising names, perceptual, functional, and encyclopaedic information about objects. Although a realist approach requires the modeller to relinquish control over the structure embedded within the corpus, the resulting structure aims to provide a closer representation of that available within the natural learning environment. Consequently, this reduces the extent to which emergent properties are determined by prior assumptions of the modeller and provides a means of examining the content of behaviour determined by naturally occurring structure within the environment. The model presented in

Rogers et al., (2004), generates the counterintuitive prediction that damaged semantic systems are more likely to perform better at specific relative to general sorting in the case of fruits. This subtle aspect of behaviour is captured as a result of the model implementing rich representations of the structure of information available within the environment.

With small corpora it is also possible to analyse the structure embedded within representations derived from natural data to identify features that may have an influence on emergent behaviour. This is demonstrated in Dilkina, McClelland and Plaut (2010), in which individual differences displayed by SD patients were modelled in an H&S framework that mapped between orthographic, action, vision and phonological layers. The behaviour of a subset of SD patients whose performance on lexical and semantic tasks did not correlate by item had been argued to result from two functionally distinct systems (e.g. Coltheart, 2004). The study demonstrated the compatibility of a single system model with the empirical data and offered an alternative explanation based on the role of spelling and concept consistency. The authors argued that observed effects emerged due to the structure embedded within representations rather than modality specific architectural constraints.

Behaviour is not only constrained by representational structure but also by the manner in which the system interacts with representations, for example the form and quantity of mappings demanded by the learning environment. H&S models have demonstrated how the framework is able to examine the consequences of such environmental factors. Dilkina et al., (2010) captures contrasts in mappings over the course of development. Training is split into two stages, with mapping from orthography to phonology only performed in the second stage. This aims to reflect the fact that learning to read only occurs at a later stage of development. The proportion and period in which certain mappings such as vision to action occur may remain relatively constant both over the course of development and populations. However, it is also true that in many cases there will be variation in the form and quantity of mapping between individuals and more broadly populations. Dilkina et al., (2008) uses this feature of the learning environment to explore one possible factor driving individual difference in SD, that being the level of prior reading experience. Within the study, prior reading experience is modelled by manipulating the amount of training on orthographic to phonological mapping. Demonstrating the influence of such factors, manipulation of this variable was able to account for four of the five SD patients behaviour. Clearly, such variation in the type of mapping performed and stage at which it's performed can have dramatic consequences for emergent properties of the system. However, predicting the nature

of such properties in complex multimodal systems is far from trivial. H&S offers a means of examining the consequences of variation in such environmental variables.

To conclude, behavioural data from the VWP suggests that language mediated visual attention is driven by the interaction of information extracted from the visual environment and speech signal at semantic, visual and phonological levels of processing. The H&S framework provides a parsimonious architecture within which the emergent properties of this complex interaction can be modelled. Previous modelling of the VWP has identified further properties of the architecture involved. These include allowing competition at multiple levels of representation, parallel activation of representations, the integration of information from multiple sources and allowing inhibitory and excitatory associations. A neural network architecture such as those used in previous implementations of the H&S framework naturally captures these characteristics.

Investigation goals

We next present a computational model of the various sources of information contributing to eye gaze in the VWP. Our aims were as follows. First, we tested whether a H&S model, with minimal computational architectural assumptions, was sufficient for replicating the effects of phonological and semantic influences on language processing in the VWP, or whether individual models combining the modal-specific features of the models of Allopenna et al., (1998) and Mirman and Magnuson (2009) would be required to effectively simulate the range of effects across these distinct modalities. Second, we tested whether the model could further generalise to simulate effects of visual information similarity in the VWP (Dahan and Tanenhaus, 2005; Huettig and Altmann, 2007). Third, we tested whether the model was further able to replicate sensitivity to the effects of presenting or not presenting the object corresponding to the target word in the various VWP experimental manipulations of visual, phonological, and semantic competitors. In each case, the model's performance is a consequence of the integrated processing of multimodal information, resulting from specified properties of the representations themselves and also the computational properties of the mappings between them.

The model we present connects visual, semantic and linguistic information to drive eye gaze behaviour. Specifically, the model was tested on its ability to replicate the following features of language mediated visual attention demonstrated in Visual World studies: (1) Fixation of onset and rhyme competitors above unrelated distractor levels in target present scenes

(Allopenna et al., 1998); (2) Fixation of visual competitors above unrelated distractor levels in both target present (Dahan and Tanenhaus, 2005) and target absent (Huettig and Altmann, 2007) scenes; and (3) Fixation of semantic competitors above unrelated distractor levels and relative to semantic relatedness in both target present (Yee & Sedivy, 2006; Mirman & Magnuson, 2009) and absent (Huettig and Altmann, 2005) scenes. We present two simulations – one with no environmental noise, and one with background environmental noise. We later show that environmental noise is necessary for replicating all aspects of behavioural data.

2.1. Modelling language-mediated visual attention in a noiseless learning environment

Method

Architecture

The architecture of the H&S neural network used within this study is displayed in Figure 2. Akin to previous H&S models it was composed of a central resource (integrative layer) consisting of 400 units that integrated modality specific information from four "visible" layers, which encoded input and output representational information. The vision layer consisted of 80 units and modelled the extraction of visual information from four spatial locations within the environment. It contained four slots each containing 20 units which extracted visual information from each of four distinct locations in the visual field. The phonological layer consisted of 60 units and encoded phonological information from the speech signal. This layer comprised six phoneme slots each represented by 10 units, such that words up to 6 phonemes in length could be represented unfolding across time. A semantic layer of 200 units represented semantic information of items, with units representing semantic features of the concept. The eye layer consisted of four units. Each unit within the eye layer was associated with one of the four locations within the model's visual field. The level of activation of an eye unit represented the probability of fixating the spatial location with which the unit was associated. All visible layers were fully connected to the central integrative layer, and the central integrative layer was in turn fully self-connected and fully connected to the eye and semantic layers.

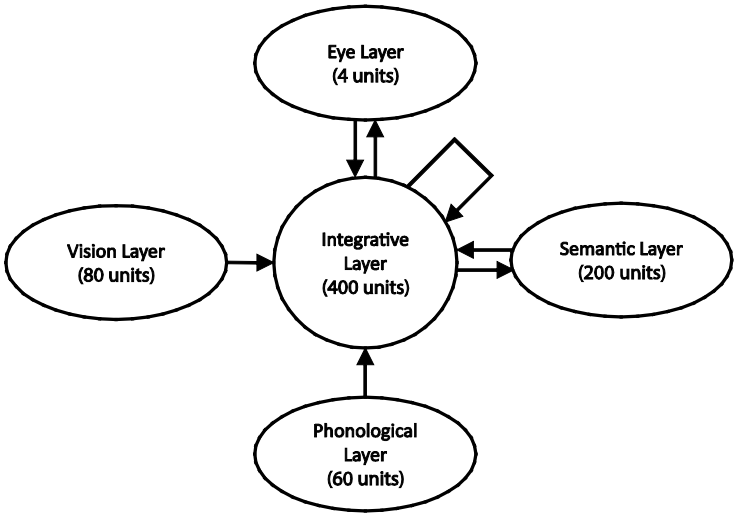

Figure 2: Network Architecture

At each time step of the model's processing, activation passed between all layers of units in the model (see appendix). During training, there were 14 time steps to enable activation to cycle between representations in the model. During testing, the number of time steps was extended to enable insight into the time-course of representational information interacting between the modalities within the model.

Artificial corpus

A fundamentalist approach (Kello & Plaut, 2000) was taken in construction of representations to ensure all aspects of the representations were controlled within simulations. Therefore, an artificial corpus composed of 200 items each with unique phonological, visual and semantic representations was constructed and used to train and test the model. Visual representations were generated to represent visual features in different spatial locations, with features representing both coarse (low frequency) and fine (high frequency) visual features. Phonological representations were encoded to create time-dependent slots for the unfolding speech, with categorical representations of phonemes shared across different words. Semantics in the model were rich, in that they were distributed feature based representations with structured relationships between items. They were also relatively sparse and discrete, reflecting behavioural studies of semantic feature-based representations (Harm & Seidenberg, 2004).

Table 1: Controls used in the construction of artificial corpora and mean cosine distance calculated between targets, competitors and unrelated items all six randomly generated corpora used to train and test models.

Modality	Item	Artificial Corpus	
		Constraint (Features shared with target)	Cosine Distance (μ, σ)
Phonological	Onset Competitor	First 3 phonemes	.259 (.026)
	Rhyme Competitor	Final 3 phonemes	.260 (.028)
	Unrelated	Max. 2 consecutive phonemes	.496 (.052)
Semantic	Near Neighbour	4 of 8 functional properties	.500 (0)
	Far Neighbour	2 of 8 functional properties	.750 (0)
	Unrelated	Max. 1 functional property	.959 (.072)
Visual	Competitor	10 of 20 visual features	.264 (.040)
	Unrelated	High and Low frequency feature vectors unique	.506 (.068)

Visual representations were encoded as 20 unit binary feature vectors, with each unit representing the presence or absence of a given visual feature. Features were assigned to items randomly with p(active) = 0.5. Phonological representations consisted of a fixed sequence of six phonemes. Words were constructed by randomly sampling phonemes from a phoneme inventory containing a total of 20 possible phonemes. Each phoneme was encoded by a 10 unit binary feature vector, with p(active) = 0.5. For semantic representations, a unique subset of 8 semantic features was randomly assigned to each item from the set of 200 possible features.

The level of overlap between items in semantic, visual and phonological dimensions was controlled (see Table 1). Within the corpus were embedded 20 target items each with either visual, near semantic, far semantic, phonological onset or rhyme competitors. Competitors were defined by the increased number of features shared with their assigned target in either a semantic, visual or phonological dimension. A consistent level of representational overlap was implemented across all modalities (other than in the case of far semantic competitors) by ensuring that the distance in terms of shared features between a target and a competitor was on average half the distance of that between a target and unrelated item in the modality that defined the competitor type. Six randomly generated corpora were generated using different

initial random seeds, to ensure that no accidental correspondences between particular representations occurred systematically.

Onset competitors shared the initial three phonemes with their corresponding target word. No two words shared their initial four phonemes. Rhyme competitors shared the final three phonemes with their assigned target. No two words shared their final four phonemes. No item within the corpus contained more than two identical phonemes per word and no more than two consecutive phonemes overlapped between two unrelated items. These constraints resulted in a cosine distance between phonological representations of 0.259 between onset competitors and targets, 0.260 between rhyme competitors and targets and 0.496 between unrelated items and targets.

The length of vectors used to encode representations in both semantic and visual dimensions was determined by the constraints placed on relationships between items in these modalities. In the case of visual competitors, 10 of 20 visual features were shared between the target and competitor with p(shared) = 1, remaining features were shared with p(shared) = 0.5. For all visually unrelated items features were shared with p(shared) = 0.5. Such controls resulted in a smaller visual feature cosine distance between visual competitors and target items than between unrelated items and targets (see Table 1).

In the semantic dimension, near semantic competitors shared 4 of 8 semantic features with their corresponding target, while 2 of 8 were shared between far semantic competitors and targets. Controls ensured that unrelated items shared a maximum of one semantic feature. Semantic feature cosine distance was least between near neighbours and targets, medial between far neighbours and targets and most between unrelated items and targets (see Table 1).

Training

Model training simulated learning experience in the natural environment that leads to the acquisition of associations between representations across modalities. We assume that individuals acquire semantic, visual and phonological knowledge of a given item through experience of repeated and simultaneous exposure to these multiple forms of representation within the natural learning environment. The model was trained on four cross modal tasks (see table 2).

To simulate the events that lead to associations between an item's visual and semantic properties, the model was trained to map from visual to semantic representations using the following procedure. An example of such an event in the natural learning environment may be viewing an item while simultaneously experiencing some aspect of its function (e.g. seeing and eating from a fork). At trial onset the model was presented with four visual representations randomly selected from the corpus assigned to the four spatial locations within the visual field. One of the four items was then randomly selected as a target and the eye unit corresponding to its location fully activated. Throughout the entire test trial small levels of variable noise were provided as input to the phonological layer to simulate ambient background sound. Once sufficient time has allowed for activation to pass from eye and visual layers to the semantic layer (at time step 3) the item's semantic representation was provided as a target and error back propagated.

Models were also trained to map between phonological and semantic representations, simulating the learning that occurs through simultaneous exposure to the sound of a given word and its semantic properties (i.e. hearing and observing the function of "fork"). First, an item was randomly selected as a target from the corpus. The phonological representation of the target was then provided to the phonological input layer as a staggered input, with one additional phoneme provided at each time step. Once activation of the fourth phoneme (uniqueness point for phoneme competitors and corresponding targets) had had sufficient time to influence activation in the semantic layer (time step 5), the item's semantic representation was provided as a target and error back propagated.

Two further tasks trained the model to orientate towards a visual representation of an item in a spatial location according to given phonological or semantic information. As previously stated we assume that in the natural learning environment individuals are repeatedly exposed simultaneously to the visual and phonological or semantic form of an item. Consequently, the learner determines the association between these representations. Mapping from phonology to location was trained by randomly selecting four items from the corpus, randomly assigning them to four locations, and randomly selecting one as the target. The visual representations relating to each of these items was presented as input to the visual layer at trial onset. At the same point in time, input of the phonological representation of the target item was initiated in the phonological layer with one additional phoneme presented per time step. Once activation relating to the fourth phoneme had had time to influence activation in the eye layer (time step

5), the eye unit corresponding to the location of the target was provided as the target and error back propagated.

For mapping from semantics to location, the trial was similar to the phonology to location task, except that all the semantic features were simultaneously activated at time step 1 and time variant noise was presented to the phonological layer for the entire training trial. Once activation from the semantic and visual layer had been provided sufficient time to influence eye layer activation (time step 2), the training signal was provided and error back propagated.

Training tasks were randomly interleaved. Within the natural learning environment we assume that individuals orientate towards or select items based on their semantic features far more frequently than they orientate towards or select items in response to hearing their name. To reflect the assumption that phonologically driven orienting occurs less frequently than semantically driven orienting, training on phonologically driven orienting was four times less likely to occur than all other training tasks.

Table 2: Temporal organisation of events in model training

Task	Vision		Phonological		Semantic		Eye	
	Description	Time step	Description	Time step	Description	Time step	Description	Time step
Visual to Semantic	4 visual representations randomly selected from corpus, 1 assigned as target	0–14	Random time invariant noise provided as input	0–14	Semantic representation of target provided post display onset	3–14	Location of target activated, all other locations inactive	0–14
Phonological to Semantic	Random time invariant noise provided as input across all 4 input slots	0–14	Speech signal of target provided as a staggered input	0–14	Semantic representation of target provided post disambiguation	5–14	No constraints on activation	
Phonological to Location	4 visual representations randomly selected from corpus, 1 assigned as target	0–14	Speech signal of target provided as a staggered input	0–14	No constraints on activation		Post disambiguation location of target activated, all other locations inactive	5–14
Semantic to Location	4 visual representations randomly selected from corpus, 1 assigned as target	0–14	Random time invariant noise provided as input	0–14	Semantic representation of target provided	0–14	Location of target activated, all other locations inactive post functional onset	2–14

All connection weights within the network were initially randomised in a uniform distribution [-0.1, 0.1]. Weights were adjusted using recurrent back-propagation with learning rate = 0.05 (see appendix). In order to simulate participants' prior ability to orientate to items based on their phonological and semantic form and identify items' semantic properties based on their

visual or phonological form, the models were required to perform with high accuracy on all four of these tasks prior to testing. To obtain this level of performance training was terminated after 1 million trials. In total 6 simulation runs of the model were trained and tested, using each of the six artificial corpora.

Results

Pre-test

Following training all models were tested to assess performance on each of the four training tasks for all items within the training corpus. Noise was presented to visual and phonological slots that did not receive target related input. For tasks presenting the target in the visual input, performance was recorded with the target tested once in each of the four locations in the visual field.

For mapping from visual to semantic representations, activation in the semantic layer was closer in terms of cosine similarity to the target item's semantic representation for all items within the training corpus. When tested on mapping from phonological to semantic representations activation in the semantic layer was also most similar to that of the target's semantic representation for all items within the training corpus.

For the phonology to location mapping task, the location of the target was selected on at least 3 of 4 test trials for 99.83% of items in the training corpus. Averaging across all phonology to location test trials the proportion of trials in which the eye unit corresponding to the location of the target was most highly activated was 92%.

For the semantics to location mapping task, the location of the target was selected on at least 3 of 4 test trials for 99.92% of items within the corpus. The overall proportion of successful semantic to location test trials was 89%.

Simulation of visual competitor effects in the VWP

To simulate the conditions under which participants were tested in Dahan and Tanenhaus (2005), the model was presented with a visual display containing a target item, a visual competitor and two unrelated distractors. Simulations of Huettig and Altmann (2007) were conducted using a similar approach yet with targets replaced by an additional distractor. In both cases, the visual input representing four items was presented at time step 0. Then onset of the phonology for the target item began at time step 5, to enable pre-processing of the

visual information. There were 480 test trials, with each item (n = 20) occurring with competitors in all possible spatial configurations (n = 24). The model's "gaze" was computed as the Luce ratio of the eye layer units, for the target, competitor, and unrelated distractor item. Figure 3 displays the performance of the model when presented with target present (Figure 3A: simulating Dahan & Tanenhaus, 2005) and target absent (Figure 3B: simulating Huettig & Altmann, 2007) scenes, averaged over each of the six simulation runs of the model. For analysis we calculated the mean fixation proportions [p(fix)] for each category of item (i.e. target, competitor or unrelated distractor) from word onset until the end of the test trial. The ratio was then calculated between the proportion of fixations towards item type A and the sum of the proportion of fixations towards item type A and B. A ratio above 0.5 would indicate greater fixation of item A.

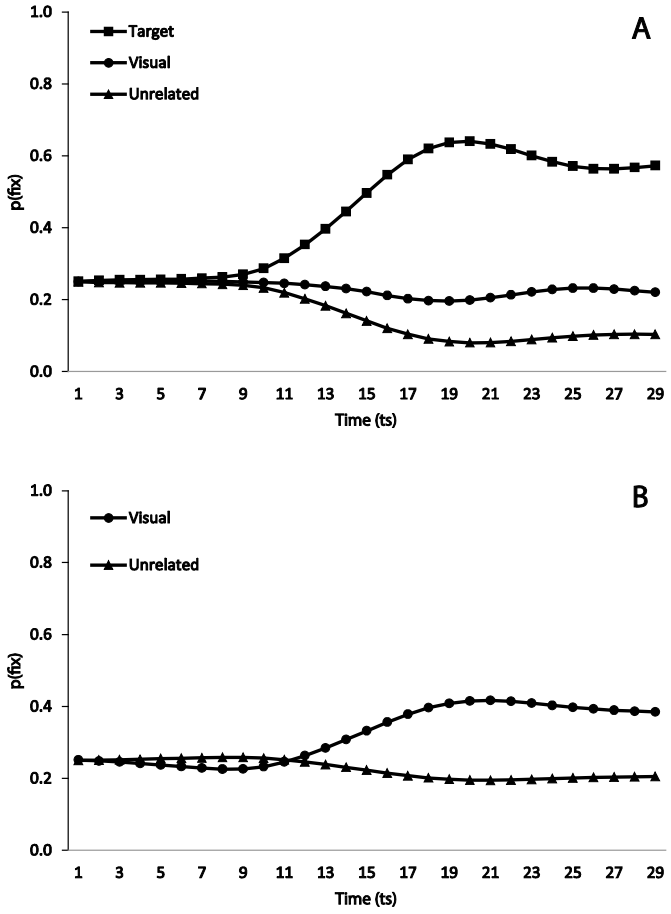

Figure 3: Proportion of fixations [p(fix)] directed toward items within scenes containing A) a target, visual competitor and two unrelated distractors B) a visual competitor and three unrelated distractors.

Although we would not anticipate substantial variation in model performance across instantiations for completeness this mean ratio (by instantiation and by item) was compared to 0.5 using one sample t-tests (cf. Dahan & Tanenhaus, 2005) to test for differences in fixation behaviour towards each category of item.

As can be observed from Figure 3, the model fixated target items [mean ratio = 0.75, $t_1(5) = 22.42$, $p < 0.001$; $t_2(19) = 78.50$, $p < 0.001$] and visual competitors [mean ratio = 0.60, $t_1(5) = 6.91$, $p = 0.001$; $t_2(19) = 18.18$, $p < 0.001$] more than unrelated distractors when scenes contained a target, visual competitor and two unrelated distractors (when by subjects and by items ratios are identical, only one ratio is presented). In target absent scenes, visual competitors were again fixated more than unrelated distractors [mean ratio = 0.58, $t_1(5) = 5.37$, $p < 0.01$; $t_2(19) = 15.290$, $p < 0.001$]. The model therefore replicates the increased fixation of visual competitors observed in Dahan & Tanenhaus (2005) and Huettig & Altmann (2007).

Simulation of semantic competitor effects in the VWP

We simulated conditions similar to those under which participants were tested in Huettig & Altmann (2005), Yee & Sedivy (2006) and Mirman & Magnuson (2009) by testing model performance when presented with displays containing a near semantic neighbour and a far semantic neighbour in addition to either the target's visual representation and a single unrelated distractor (Figure 4A: simulating Mirman & Magnuson, 2009 and Yee & Sedivy, 2006) or two unrelated distractors (Figure 4B: Simulating Huettig & Altmann, 2005). As for the visual competitor effects, all items were presented in all combinations of positions in the visual input (480 trials in total), and again pre-processing of the visual features of the display were enabled by commencing word onset after a short delay (time step 5). Figure 4 presents the average fixation proportions over time displayed by the model towards each category of item presented in both test conditions.

In target present trials, targets [mean ratio = 0.75, $t_1(5) = 25.89$, $p < 0.001$; $t_2(19) = 79.61$, $p < 0.001$], near semantic neighbours [mean ratio = 0.58, $t_1(5) = 5.37$, $p < 0.01$; mean ratio = 0.57, $t_2(19) = 9.89$, $p < 0.001$] and far semantic neighbours [mean ratio = 0.52, $t_1(5) = 2.82$, $p < 0.05$; mean ratio = 0.51, $t_2(19) = 4.07$, $p < 0.01$] were all fixated more than unrelated distractors. A similar pattern of behaviour was observed when the model was tested on target absent trials, with both near [mean ratio = 0.58, $t_1(5) = 6.30$, $p < 0.01$; mean ratio = 0.57, $t_2(19) = 10.67$, $p < 0.001$] and far semantic neighbours [mean ratio = 0.53, $t_1(5) = 1.80$, $p >$

0.1; mean ratio = 0.52, $t_2(19) = 7.04$, $p < 0.001$] fixated more than unrelated items. Also in-line with behavioural findings far semantic neighbours were fixated less than near semantic neighbours, in both target absent [mean ratio = 0.44, $t_1(5) = -3.36$, $p < 0.05$; mean ratio = 0.45, $t_2(19) = -8.13$, $p < 0.01$] and target present [mean ratio = 0.44, $t_1(5) = -3.36$, $p < 0.05$; mean ratio = 0.44, $t_2(19) = -6.97$, $p < 0.001$] conditions. The model therefore replicates the increased fixation of semantic competitors in both target absent and target present scenes as observed by Huettig & Altmann (2005) and Yee & Sedivy (2006) respectively, in addition to the graded effect of semantic similarity as reported in Mirman and Magnuson (2005).

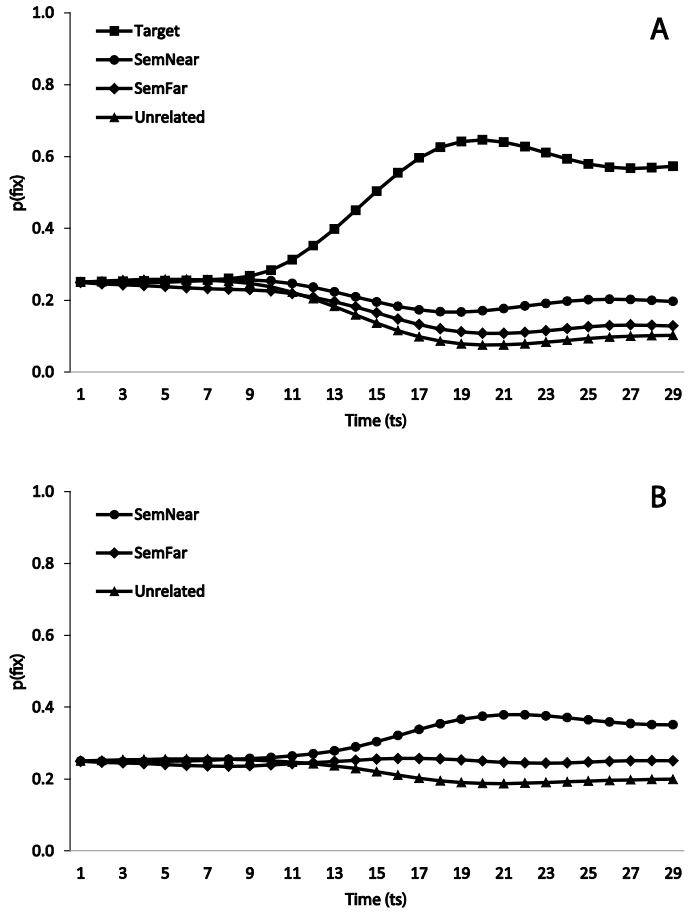

Figure 4: Proportion of fixations [p(fix)] directed toward items within scenes containing A) a target, a near semantic neighbour (SemNear), a far semantic neighbour (SemFar) and an unrelated distractor, B) a near semantic neighbour (SemNear), a far semantic neighbour (SemFar) and two unrelated distractors.

Simulation of phonological competitor effects in the VWP

To simulate the conditions under which participants were tested in Allopenna et al.'s (1998) study, the model was presented with scenes containing visual representations of a target item

in addition to an onset competitor, a rhyme competitor and an unrelated distractor. For completeness we also tested model performance in a target absent condition (i.e. scenes containing onset competitor, rhyme competitor and two unrelated distractors). In every other way, simulations were conducted exactly as for the visual and semantic competitor simulations. Figure 5 shows the average fixation proportions over time displayed by the model towards each category of item in test displays in both target present (Figure 5A) and target absent (Figure 5B) conditions.

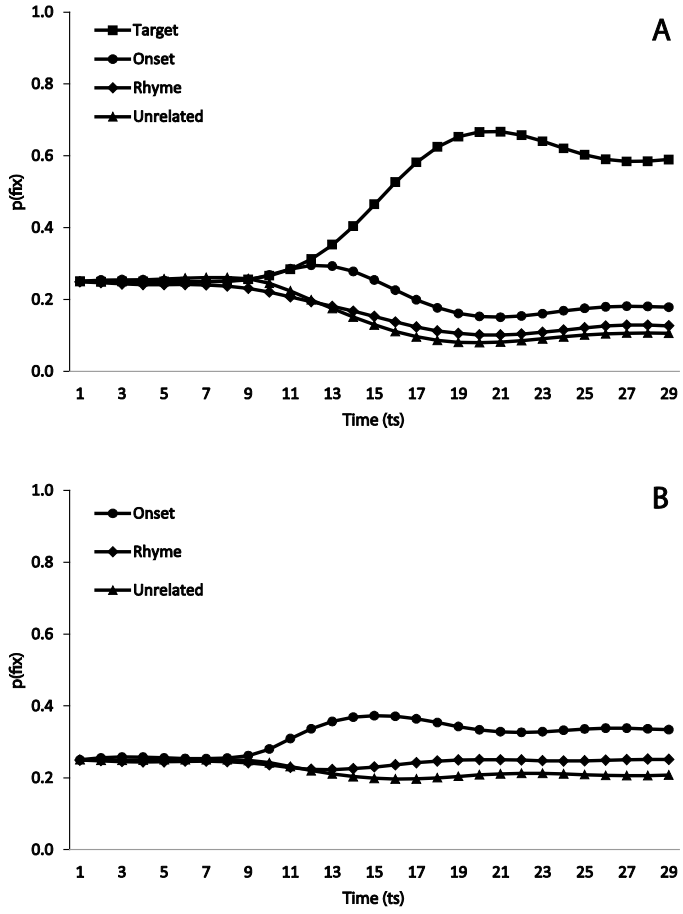

Figure 5: Proportion of fixations [p(fix)] directed toward items within scenes containing A) a target, an onset competitor, a rhyme competitor and an unrelated distractor, B) an onset competitor, a rhyme competitor and two unrelated distractors.

In target present trials, target items [mean ratio = 0.75, $t_1(5) = 26.06$, $p < 0.001$, $t_2(19) = 66.45$, $p < 0.001$] and onset competitors [mean ratio = 0.58, $t_1(5) = 6.20$, $p < 0.01$, $t_2(19) = 16.52$, $p < 0.001$] were fixated more than unrelated distractors. However, the model fixated rhyme competitors at levels similar to unrelated distractors [mean ratio = 0.51, $t_1(5) = 1.75$, $p > 0.1$, $t_2(19) = 1.69$, $p > 0.1$]. On target absent trials both onset [mean ratio = 0.59, $t_1(5) =$

8.29, $p < 0.001$; $t_2(19) = 15.62$, $p < 0.001$] and rhyme [mean ratio = 0.53, $t_1(5) = 5.62$, $p < 0.01$; mean ratio = 0.52, $t_2(19) = 3.05$, $p < 0.01$] competitors were fixated more than unrelated items. Allopenna et al., (1998) observed increased fixation of both onset and rhyme competitors in target present scenes. Model performance replicated the increased fixation of onset competitors displayed by participants. The model also displayed increased fixation of rhyme competitors although this effect was only clearly observable on target absent trials.

Discussion

The model was able to replicate a broad range of single modality word level effects described in the visual world literature, using a single architecture, and incorporating a single shared resource mapping between the modalities. The network replicates findings displaying a bias toward fixating items that overlap with spoken target words in either a visual, semantic or phonological dimension in both target present and absent scenes.

Importantly, the model captures differences in the effect of representational overlap between modalities. The model displays a graded effect of semantic overlap with the probability of fixating semantically related items proportional to the number of semantic features shared between the target and competitor. In a departure from the procedure used in Mirman & Magnuson (2009), within the above simulations both near and far semantic competitors were presented within the same display. Our simulations indicate that far semantic neighbour effects are robust to the additional competition that may result from the presence of closer semantic neighbours within the same scene.

For phonological overlap, the effect was dependent on the temporal location of overlapping features within the representation. Phonological overlap in onsets had a greater influence on fixation behaviour than in rhymes, with the latter resulting in only marginal effects of overlap. Although many studies have demonstrated their existence (see Allopenna et al., 1998; Desroches et al., 2006; McQueen & Viebahn, 2007; McQueen & Huettig, 2012), rhyme effects tend to be weak and less robust than onset effects. However, a recent study by McQueen and Huettig (2012) provides evidence that the comparative onset effect is modulated by the level of noise present in the speech signal. They argue that the presence of noise influences the weight placed on initial phonemes as a predictor of the intended word. For example, in a noisy environment sounds heard may not necessarily relate to the identity of the target. Therefore, to make a judgement regarding an item's identity the system benefits from examining evidence from a larger portion of the auditory signal. This work highlights a

weakness of current model training and testing, in that the model's learning environment always provided perfect perceptual input of an item in both visual and phonological representations. In the natural learning environment in which participants acquire their knowledge of items, the cognitive system is frequently receiving impoverished representations. This is particularly true in the case of speech, in which factors such as background noise or between speaker variation means that the speech signal received is likely to resemble only a very noisy version of the canonical form. The following simulations extend the model by adding noise to the phonological representations to which the model is exposed during training.

2.2. Modelling language-mediated visual attention in a noisy learning environment

Method

To simulate exposure to noisy phonological input in the natural learning environment, the simulations were repeated but with noise applied to the phonological input during the training stage only. Noise was implemented by randomly switching the binary value of each unit within the phonological representation with p=0.2. Noise was randomly generated for each training trial. To ensure comparable levels of performance between fully trained models on all four training tasks, the number of training trials performed was increased by 50%. In all other respects the procedure used to train and test the noisy model was identical to that applied to the previously detailed noiseless model.

Results

Pre-test

The noisy model displayed the same high level of performance on both visual to semantic mappings and phonological to semantic mappings as displayed by the noisless model. In both cases, the noisy model produced activation in the semantic layer most similar (cosine similarity) to the target item's semantic representation for all items within the training corpus.

Performance on orientation tasks was also similar for models trained in both noise conditions. On phonological orienting test trials, the noisy model selected the location of the target on at least 3 of 4 test trials for 99.75% of items in the training corpus. The overall proportion of

correct phonological orienting test trials (trials in which the eye unit corresponding to the location of the target was most highly activated) was 87% for the noisy model. When comparing the proportion of correct trials across instantiations between noise conditions, noiseless models performed significantly better than noisy models on this task ($p=0.01$).

Noisy models correctly selected the location of the target as indicated by the presence of its semantic representation on at least 3 of 4 test trials for all items within the corpus. Overall accuracy on semantic orienting tasks for the noisy model was 90% ($\sigma = 0.02$). The difference between noisy and noiseless models was not significant on this task when comparing across instantiations.

Simulation of visual competitor effects

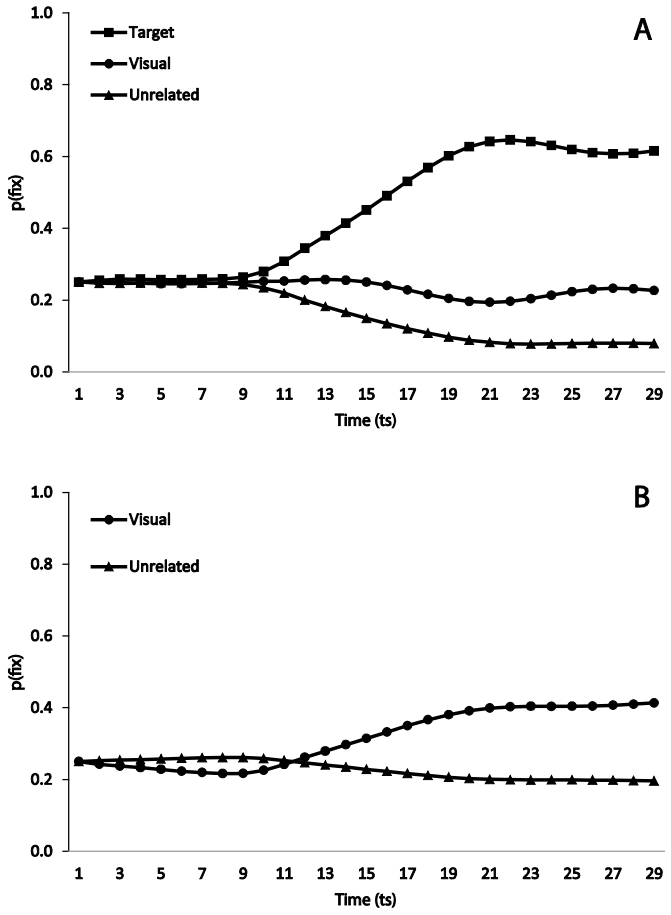

Figure 6: Proportion of fixations [p(fix)] directed toward items within scenes containing A) a target, visual competitor and two unrelated distractors B) a visual competitor and three unrelated distractors; by the model trained in a noisy learning environment.

Figure 6 displays the performance of the noisy model when tested on scenes containing a visual competitor in addition to either the visual representation of the target and two unrelated distractors (Figure 6A) or no target and three unrelated distractors (Figure 6B).

On target present trials, both the targets [mean ratio = 0.77, $t_1(5) = 27.21$, $p < 0.001$; $t_2(19) = 89.97$, $p < 0.001$] and visual competitors [mean ratio = 0.62, $t_1(5) = 7.60$, $p < 0.01$; $t_2(19) = 22.22$, $p < 0.001$] were fixated more than unrelated distractors. Visual competitors were also fixated above distractor levels on target absent trials [mean ratio = 0.60, $t_1(5) = 14.52$, $p < 0.001$; mean ratio = 0.59, $t_2(19) = 18.75$, $p < 0.001$].

Simulation of semantic competitor effects

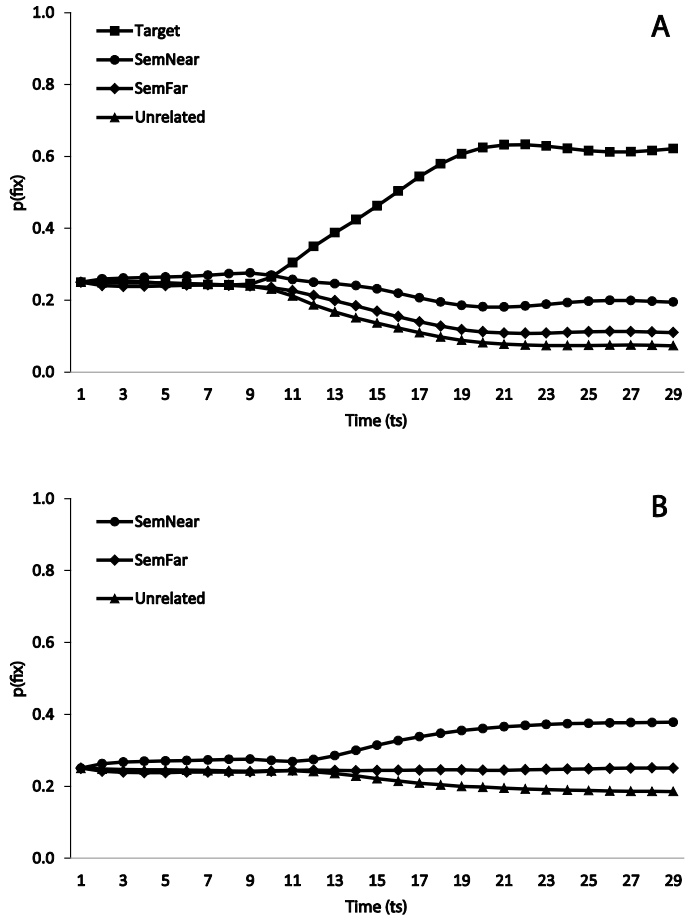

Figure 7: Proportion of fixations [p(fix)] directed toward items within scenes containing A) a target, a near semantic neighbour (SemNear), a far semantic neighbour (SemFar) and an unrelated distractor, B) a near semantic neighbour (SemNear), a far semantic neighbour (SemFar) and two unrelated distractors; by the model trained in a noisy learning environment.

The fixation behaviour displayed by the noisy model on trials containing semantic competitors can be seen in Figure 7. The model was tested on scenes containing a near and far semantic neighbour in addition to either the target and a single unrelated distractor (Figure 7A) or no target and two unrelated distractors (Figure 7B).

On target present trials, targets [mean ratio = 0.78, $t_1(5) = 29.48$, $p < 0.001$; mean ratio = 0.76, $t_2(19) = 102.21$, $p < 0.001$], near semantic neighbours [mean ratio = 0.62, $t_1(5) = 6.42$, $p < 0.01$; mean ratio = 0.60, $t_2(19) = 18.389$, $p < 0.001$] and far semantic neighbours [mean ratio = 0.54, $t_1(5) = 2.31$, $p < 0.1$; mean ratio = 0.52, $t_2(19) = 5.934$, $p < 0.001$] were all fixated more than unrelated distracters. On target absent trials, both near [mean ratio =0.60, $t_1(5) = 13.78$, $p < 0.001$; mean ratio = 0.59, $t_2(19) = 22.51$, $p < 0.001$] and far [mean ratio = 0.53, $t_1(5) = 2.75$, $p < 0.05$; mean ratio = 0.52, $t_2(19) = 7.13$, $p < 0.001$] semantic neighbours were again more likely to be fixated than unrelated items. When comparing between near and far semantic competitors, far neighbours were fixated less than near neighbours both in target present [mean ratio = 0.42, $t_1(5) = -12.45$, $p < 0.001$; $t_2(19) = -12.81$, $p < 0.001$] and absent [mean ratio = 0.43, $t_1(5) = -11.81$, $p < 0.001$; $t_2(19) = -15.84$, $p < 0.001$] trials.

Simulation of phonological competitor effects

Finally, the model was tested on scenes containing onset and rhyme competitors in addition to either the target and a single unrelated distractor (Figure 8A) or two unrelated distractors (Figure 8B).

In target present scenes, the model displayed increased fixation of target items [mean ratio = 0.77, $t_1(5) = 36.71$, $p < 0.001$; $t_2(19) = 76.149$, $p < 0.001$], onset competitors [mean ratio = 0.60, $t_1(5) = 6.51$, $p < 0.01$; mean ratio = 0.61, $t_2(19) = 18.11$, $p < 0.001$] and rhyme competitors [mean ratio = 0.54, $t_1(5) = 3.13$, $p < 0.05$; $t_2(19) = 6.842$, $p < 0.001$] in comparison to unrelated distractors. Onset [mean ratio = 0.60, $t_1(5) = 11.09$, $p < 0.001$; $t_2(19) = 17.35$, $p < 0.001$] and rhyme competitors [mean ratio = 0.54, $t_1(5) = 3.13$, $p <0.05$; $t_2(19) = 8.90$, $p < 0.001$] were also fixated more than distractors in target absent scenes.

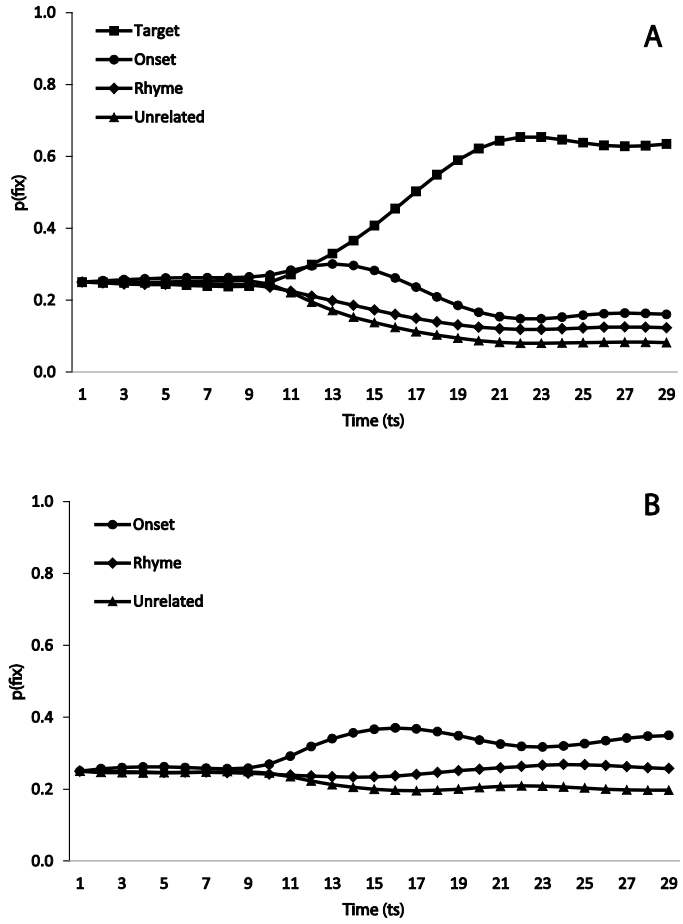

Figure 8: Proportion of fixations [p(fix)] directed toward items within scenes containing A) a target, an onset competitor, a rhyme competitor and an unrelated distractor, B) an onset competitor, a rhyme competitor and two unrelated distractors; by the model trained in a noisy learning environment.

Discussion

The above results demonstrate that the model of language-mediated visual attention presented in this paper is still able to replicate a broad range of features of language-mediated visual attention when trained in a noisy learning environment. Further, and as predicted, by representing noise in the speech signal during training, we are able to replicate additional features of language-mediated visual attention, specifically sensitivity to rhyme competitors.

Table 3: Table comparing the results of both noiseless and noisy simulations with behavioural results reported in the VWP literature. The items displayed within scenes in each empirical study are listed with observed competitor effects highlighted in bold. Competitor-Distractor ratios (by subject/instantiation) in parentheses if reported; √ = behavioural effect

*replicated; X = failure to replicate behavioural effect; * = Study presented near and far semantic competitors on separate trials.*

Study		Scene				Effect A			Effect B		
Author	Year	Item 1	Item 2	Item 3	Item 4	Behav.	Noiseless	Noisy	Behav.	Noiseless	Noisy
Allopenna et al.	1998	Target	Onset (A)	Rhyme (B)	Dist	√	√ (.58)	√ (.60)	√	X (.51)	√ (.54)
Dahan & Tanenhaus	2005	Target	Visual (A)	Dist	Dist	√ (.7)	√ (.60)	√ (.62)			
Huettig & Altmann	2007	Visual (A)	Dist	Dist	Dist	√	√ (.58)	√ (.60)			
Yee & Sedivy	2006	Target	Sem (A)	Dist	Dist	√	√ (.58)	√ (.62)			
Huettig & Altmann	2005	Sem (A)	Dist	Dist	Dist	√	√ (.58)	√ (.60)			
Mirman & Magnuson*	2009	Target	Near Sem (A)	Far Sem (B)	Dist	√	√ (.58)	√ (.62)	√	√ (.52)	√ (0.54)

3. General discussion

The multimodal integration model (MIM) presented here offers a description of the information and processes underlying language mediated visual attention and a potential explanation for how it is acquired. The model accomplishes these effects with minimal imposed constraints on information processing modules or channels, and performance in the model is thus driven by representational structure and the different requirements of forming mappings between the distinct types of information. Language mediated visual attention is simulated as a function of the integration of past and current exposure to visual, linguistic and semantic forms. The model thereby provides an explicit description of the connection between the modality-specific input from language and vision and the distribution of eye gaze in language mediated visual attention.

The model replicated the following features of language mediated visual attention demonstrated in VWP studies: Fixation of onset and rhyme competitors above unrelated distractor levels in target present scenes (Allopenna et al., 1998); (2) Fixation of visual competitors above unrelated distractor levels in target present (Dahan & Tanenhaus, 2005) and target absent (Huettig & Altmann, 2007) scenes; and (3) Fixation of semantic competitors above unrelated distractor levels and relative to semantic relatedness in both target present (Yee & Sedivy, 2006; Mirman & Magnuson, 2009) and absent (Huettig & Altmann, 2005) scenes. A summary of the effects replicated by the model is presented in Table 3.

The results of the above simulations met the objectives of our study as follows. First, the model demonstrates that a H&S model, with minimal computational architectural assumptions, was sufficient for replicating the word level effects of phonological and semantic influences on language processing in the VWP. The simulation results replicate a

broad range of the word level effects described within the VWP literature as features of this complex cognitive ability, without requiring separate resources or individually trained pathways between distinct representational information. Second, the model further generalised to replicate the effects of visual similarity in the VWP and sensitivity to the effects of presenting or not presenting the target object in various experimental manipulations of visual, phonological and semantic competitors.

Within our model language mediated visual attention is described as an emergent property of the structure of representations present in the natural environment and the task demands imposed on the system by that environment. Knowledge of an item is acquired by repeated, simultaneous exposure to its multiple forms. For example, hearing the name of an object while looking at it, or experiencing the function of an item while hearing its name. Such experience leads to associations between the properties defining an object in separate modalities. With repeated and simultaneous exposure to their various forms inhibitory or excitatory connections between such properties are strengthened in order for the system to efficiently map between representations or carry out a given task. In this way, the model provides an explicit and detailed description of how multimodal knowledge of an item is acquired and stored, in addition to how complex multimodal behaviours such as selecting an item based on its function may be achieved and acquired. Thus, the model argues that many word level features of language mediated visual attention are a necessary consequence of developing multimodal knowledge of items through such a mechanism.

Critically, the model captures contrasts in the effect of overlap in differing modalities. For example, for items that only overlap in a semantic dimension the probability of the model fixating an item is directly proportional to the number of semantic features the two items share. This replicates findings observed in the VWP in which the probability of fixating items has been predicted by semantic norming data (Mirman & Magnuson, 2009) and corpus-based measures of semantic similarity (Huettig et al., 2006). However, in the case of phonological overlap, the temporal location of the overlapping phonemes has a critical influence on the resulting effect. The model replicates the effects of phonological overlap observed in Allopenna et al., (1998) with items that share initial phonemes fixated earlier and with greater probability than items that share phonemes in final positions.

Within the model the level of overlap between target and competitor was strictly controlled both across modalities and between rhyme and onset competitors. Contrasts in fixation

behaviour toward differing categories of competitor therefore arise as an emergent property of differences in the structural characteristics of representations in each modality. For example, speech unfolds over time. Therefore, phonological representations have a temporal, sequential component not possessed by semantic or visual representations. As the speech signal gradually manifests, early phonemes provide a good, or in the case of a noiseless learning environment they provide a perfect, predictor of the intended word. Therefore, any item that shares the same initial sequence of phonemes with the target is more likely to be fixated by the model. By the time later phonemes are available, the system already has sufficient information, in the case of the noiseless simulations, to identify the target and therefore information provided by later phonemes does not have the opportunity to exert influence on target selection. It is for this reason increased sensitivity to rhyme competitors is displayed by a model trained in a noisy environment compared to one trained in a noiseless environment in which onset phonemes are perfect predictors of the unfolding word. Behaviour of the noisy model demonstrates that introducing a low level of noise to speech in the learning environment is sufficient to allow the subtle influence of rhyme overlap to emerge.

This line of argument overlaps with the explanation provided in Magnuson et al., (2003) for the observed reduced sensitivity over the course of word learning to rhyme competitors. They argue that it takes time for the system to learn the value of early phonemes as predictors of the unfolding word. Therefore, at earlier stages of development other overlapping aspects of a word's phonology may exert equal or greater influence on target selection. In a noiseless environment an optimal model should display no influence of rhyme overlap, as sufficient information is carried by initial phonemes to correctly identify the target item. However, in a noisy environment the optimal model would display sensitivity to rhyme overlap proportional to the level of noise in the environment, as this will dictate the probability that the rhyme competitor is the true target given the initially perceived input. Given this line of argument, it is not only external noise that would dictate a system's sensitivity to rhyme overlap but also the level of noise or error within the system itself. For example, noise simulated within the current model could equally reflect errors in phonological perception or fluctuations in attention, the contribution of which could possibly be examined through further combined modelling and VWP studies.

Similar to TRACE (McClelland & Elman, 1986), our model displays sensitivity to overlap in both phonological onsets and rhyme. However, there are differences between the models in

their explanation for these effects. As in the model we present, TRACE is able to exploit similarity at all points within the phonological form of the word in terms of co-activating phonological competitors. However, unlike some previous models (Norris, 1994; Marslen-Wilson, 1987, 1993; Magnuson et al., 2003) and the model presented in this paper the disparity between sensitivity to cohort and rhyme competitors in TRACE is not driven by bottom-up mismatch but instead purely by onset competitors accumulating activation prior to rhyme competitors due to their inherent temporal advantage (Magnuson et al., 2003).

Many similarities are shared between our computational model and the theoretical model of language mediated visual attention proposed in Huettig & McQueen (2007). Both models argue that behaviour in the VWP is driven by matches between information extracted from visual and auditory input at phonological, semantic and visual processing levels. However, they differ subtly in how this is implemented. Huettig and McQueen suggest that contrasts in fixation dynamics displayed towards each category of competitor are driven by aspects of the systems architecture, specifically temporal contrasts in the nature of the cascade of information between modalities. For example, they argue that early fixation of phonological competitors reflects earlier activation of phonological representations in the speech-recognition system, with activation then later cascading to semantic and visual levels of processing, which in turn leads to the later increased fixation of visual and semantic competitors. In contrast, in the model proposed in the current paper, eye gaze is a continuous measure of the simultaneous integration of information activated across all three modalities. Therefore, activation of an item's phonological representation cannot influence gaze independent of currently activated visual and semantic representations.

Huettig & McQueen (2007) highlight the value of the VWP as a tool for probing finer aspects of the architecture of the cognitive system, as eye gaze offers a fine grained measure of the information activated over time. By combining this rich behavioural measure with the current model it may be possible to further examine more subtle aspects of the systems architecture that have so far proved difficult to isolate without implementation. We hope to test whether the parsimonious architecture presented in this paper is compatible with the data provided by Huettig & McQueen (2007). It remains to be seen whether such an architecture can also offer explanation for the complex time course dynamics that emerge when competitors from multiple modalities are presented simultaneously within the same display. The results of our simulations establish the applicability of the shared resource model to account for interactions between pairs of modalities. We demonstrate its ability to replicate a range of effects

involving visual-semantic and visual-phonological interactions (see Table 3), a necessary precursor before extending to multiple interactive effects.

Within the model we present, noise is only applied to phonological input. However, in the human cognitive system, perceptual input from all modalities provides only a noisy representation of the true nature of objects in the environment. It may therefore be interesting to also extend the model to capture environmental noise in visual input. Unlike speech, visual descriptions of objects can often be improved by gathering additional information regarding its visual features over time. The literature indicates that certain groups of visual features are activated earlier than others, for example low spatial frequency information has been shown to be recruited early and rapidly by the visual system (Bar, 2003). A detailed implementation of such features of visual processing is yet to be implemented within the model. It is possible that such features may have interesting consequences for language mediated visual attention. The model described in this paper potentially provides a means of exploring such questions.

Further applications of the model can be found in on-going experimental work that suggests that the relative influence of representational overlap in semantic, visual and phonological dimensions fluctuates over the course of child development (Mani & Huettig, in preparation). As previously discussed, model training simulates the interactions between the cognitive system and the learning environment through which the system acquires knowledge of objects in the world. Through sampling performance of the model as it moves through the training process it is possible to extract measures of its behaviour on individual tasks across the course of development. It may therefore be possible, in this way, to explore the developmental story of language mediated visual attention and provide an explicit description of the mechanism driving observed variation across development.

The model also provides scope for modelling individual differences in language mediated visual attention observed between mature populations. In a recent study conducted by Huettig, Singh and Mishra (2011), language mediated visual attention varied as a consequence of literacy training. Their results showed that whereas a high-literate population demonstrated phonological competitor effects similar to those previously discovered (Allopenna et al., 1998), low-literates' eye gaze did not display sensitivity to phonological overlap between spoken target words and items presented in a visual display. Instead low-literates' gaze was strongly influenced by semantic relationships between items. One explanation for this difference that could be tested in the current model is whether observed

differences in language mediated visual attention between low and high literates emerge a consequence of finer grained processing of the speech signal that follows from increased literacy training (cf. Ziegler & Goswami, 2005). The modelling framework presented in this paper allows manipulation of environmental variables such as the form of representations processed and the tasks performed in the learning environment. By manipulating such variables, it becomes possible to test theoretical explanations for these observed individual differences (see Smith, Monaghan & Huettig, 2013).

As in previous H&S models, emergent properties of this style of model are dictated by multiple factors including environmental variables such as the structure of representations and the type and frequency of mappings performed, in addition to resource-related factors such as the number of units within the central resource. With so many degrees of freedom open to the modeller with which to fit H&S models to data sets, it is crucial that steps are taken to avoid simply data fitting and instead develop a model able to probe important theoretical questions (see Seidenberg & Plaut, 2006). Any assumptions made in the model development process should be justifiable with clear theoretical motivation. One effective method of model validation is to extract from a model testable non-trivial predictions. Our model of VWP effects was effective in simulating a broad range of behaviour using a single set of parameters. When noise was present in the training environment, we effectively simulated processing of visual, phonological and semantic competitors and in differing situations – when targets were present or absent from the visual input to the model. Furthermore, subtle patterns of fixations over time were demonstrated by the model that were similar to behavioural data. Figure 1 illustrated the effect of semantic competitors in behavioural data, with an emerging preference for the target, and a later, but smaller, diverging effect of near and distant semantic competitors. A similar pattern is illustrated in the model, as shown in Figure 4. Data-fitting to such nuanced patterns of behaviour is likely to require many free parameters, and so our model's dynamics are effective in generalising to a broad range of behavioural effects.

Connecting modalities via a central resource as in H&S does not provide the only solution for connecting the various modalities known to play a role in language mediated visual attention. Other models are possible in which one builds in additional modalities separately. The advantage of the model presented in this paper is that the effects reported are emergent. A critical feature of the model's architecture is the shared resource that intervenes between modal-specific representational systems. Such an architecture is characteristic of H&S

models (Dilkina et al., 2010; Plaut, 2002; Rogers et al., 2004), and ensures that no unnecessary assumptions about how mappings are formed between distinct representations are included in the model. Furthermore, the shared resource in the model appears to parsimoniously support the interactions between multiple representations that are so characteristic of complex language-processing behaviour. Processing in each of the modalities described within the model is likely to involve complex hierarchical systems (see Simmons & Barsalou, 2003; McNorgan, Reid & McRae, 2011). However, the results of our study demonstrate that a parsimonious H&S architecture is able to capture a diverse range of effects reported in the language mediated visual attention literature (see Table 3). We argue that the model presented operates at a suitable level of abstraction to act as a meaningful proxy for the cognitive system that supports language mediated visual attention. In doing so the model provides a valuable contribution in describing the nature of the representations and processes involved in this complex multimodal behaviour performed by individuals on a daily basis, and further it offers a tool through which the factors driving individual differences in language mediated visual attention can be examined.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, *38(4),* 419-439.

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta psychologica*, *137(2),* 181-189.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15(4),* 600-609.

Barsalou, L. W., Kyle Simmons, W., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, *7(2),* 84-91.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge and London: MIT Press.

Coltheart, M. (2004). Are there lexicons?. *The Quarterly Journal of Experimental Psychology Section A, 57(7),* 1153-1171.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6(1),* 84-107.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science, 23(3),* 371-414.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology-General, 132(2),* 163-200.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic bulletin & review, 12(3),* 453-459.

Desroches, A. S., Joanisse, M. F., & Robertson, E. K. (2006). Specific phonological impairments in dyslexia revealed by eyetracking. *Cognition*, *100(3),* B32-B42.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, *25(2),* 136-164.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010). Are there mental lexicons? The role of semantics in lexical decision. *Brain research*, *1365,* 66-81.

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, *110(3),* 412-431.

Elman, J. L. (1990). Finding structure in time. *Cognitive science, 14(2),* 179-211.

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, *65,* 231–262.

Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. C*ognitive Psychology, 53*(4), 310-344.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological review*, *111(3),* 662-720.

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96(1),* B23-B32.

Huettig, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica, 121(1),* 65-80.

Huettig, F., & Altmann, G. T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, *15(8),* 985-1018.

Huettig, F., Olivers, C. N., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica*, *137(2),* 138-150.

Huettig, F., Mishra, R. K., & Olivers, C. N. (2011). Mechanisms and representations of language-mediated visual attention. *Frontiers in psychology, 2,* 394.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, *137(2),* 151-171.

Kello, C. T., & Plaut, D. C. (2000). Strategic control in word reading: Evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology Learning Memory and Cognition*, *26(3),* 719-750.

Kintsch, W. (2008). Symbol systems and perceptual representations. *Symbols and embodiment: Debates on meaning and cognition*, 145-163.

Kukona, A., & Tabor, W. (2011). Impulse processing: A dynamical systems model of incremental eye movements in the visual world paradigm. *Cognitive science*, *35(6),* 1009-1051.

Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and differentiation in semantic memory. *Annals of the New York Academy of Sciences*, *1124(1),* 61-76.

Lambon Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences*, *107(6),* 2717-2722.

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in Psychology, 3,* 54.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: studies with artificial lexicons. *Journal of Experimental Psychology: General, 132(2),* 202-227.

Mahon, B.Z. & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology*, *102*, 59-70.

Mani, N., Johnson, E., McQueen, J. M., & Huettig, F. (2013). How yellow is your banana? Toddlers' language-mediated visual search in referent-present tasks. *Developmental Psychology, 49(6),* 1036-1044.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua, 92*, 199-227.

Mayberry, M. R., Crocker, M. W., & Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive science*, *33(3),* 449-496.

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *The Quarterly Journal of Experimental Psychology*, *60(5),* 661-671.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America*, *131*, 509-517.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18(1),* 1-86.

McNorgan, C., Reid, J., & McRae, K. (2011). Integrating conceptual knowledge within and across representational modalities. *Cognition, 118(2),* 211-233.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & cognition*, *37(7),* 1026-1039.

Monaghan, P. & Nazir, T. (2009). Modelling sensory integration and embodied cognition in a model of word recognition. In J. Mayor, N. Ruh, & K. Plunkett (Eds.), *Connectionist models of behaviour and cognition II.*, pp.337-348. Singapore: World Scientific.

Monaghan, P. & Mattock, K. (2012). Integrating constraints for learning word-referent mappings. *Cognition, 123*, 133-143.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, 1(2), 263–269.

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology*, *19(7),* 603-639.

Pobric, G., Jefferies, E., & Ralph, M. A. L. (2007). Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. *Proceedings of the National Academy of Sciences*, *104(50),* 20137-20141.

Prinz, J. J. (2002). *Furnishing the mind: concepts and their perceptual basis*. Cambridge, MA: MIT Press.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review*, *111(1),* 205-234.

Seidenberg, M. S., & Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. *From inkmarks to ideas: Current issues in lexical processing*, 25-49.

Simmons, W. K., & Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology*, *20(3-6),* 451-486.

Smith, A. C., Monaghan, P., & Huettig, F. (2013). Modelling the effects of formal literacy training on language mediated visual attention. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (2013). Austin, TX: Cognitive Science Society.

Spivey, M. (2008). *The continuity of mind* (Vol. 40). Oxford University Press, USA.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, *268(5217),* 1632-1634.

Vandenberghe, R., Price, C., Wise, R., Josephs, O., & Frackowiak, R. S. (1996). Functional anatomy of a common semantic system for words and pictures. *Nature*, *383(6597),* 254-6.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal Of Experimental Psychology Learning Memory And Cognition, 32(1),* 1-14.

Yoon, E.Y., Heinke, D., & Humphreys, G.W. (2002). Modelling direct perceptual constraints on action selection: The Naming and Action Model (NAM). *Visual Cognition*, *9*, 615-661.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological bulletin*, *131(1),* 3-29.

Appendix

Neural networks simulations were conducted using Mikenet version 8.0 developed by M. W. Harm (www.cnbc.cmu.edu/~mharm/research/tools/mikenet/), a collection of libraries written in the C programming language for implementing and training connectionist networks.

Networks were trained using the continuous recurrent backpropagation through time training algorithm provided in Mikenet (crbp.c) which implements Pearlmutter (1989). Unit activation was calculated using a logistic activation function and sum squared error was used to calculate error. Time within the network was modelled by using 14 samples and an integration constant of 0.25. All other parameters were set to the default values implemented in Mikenet version 8.0.

Chapter 3

Connecting language and vision: Examining the development and internal processing of a multimodal integration model (MIM) of language-mediated visual attention[1]

Abstract

Language is generally processed in rich contexts, with speech accompanied by substantial visual and contextual information. Language mediated visual attention studies demonstrate the subtle time-course of information integration from multiple modalities during language processing, typically with auditory information influencing processing before semantic and visual information affect behaviour. We present a computational model of multimodal integration for language processing, based on the hub-and-spoke model of semantic processing but applied to multiple modalities. The model was trained to learn associations between pairs of modalities – auditory, visual, and semantic. Assessing the model as it learned reflected the developmental trajectory of children's performance in language processing when multiple modalities are present. Furthermore, the model simulated the subtle time-course of language mediated visual attention studies, which were a consequence of the computational properties of stimuli from different modalities, and not due to architectural constraints on information integration. In particular, the model shows that behaviour is consistent with a continuous process in which multiple information types interact in parallel. The model provides a challenge to descriptive theoretical accounts proposing that multimodal language processing depends on modular, cascading, information processing constraining integration of information between modalities.

[1] *Adapted from Smith, A. C., Monaghan, P., & Huettig, F. (in preparation). Connecting language and vision: A multimodal integration model (MIM) of language mediated visual attention.*

1. Introduction

Understanding language processing requires determining which of the multiple information sources in the environment the individual brings to bear on the task (Hollich et al., 2000; Moore, Angelopoulos, & Bennett, 1999; Yurovsky, Boyer, Smith, & Yu, 2013). There has been a recent growth in studies of the way in which multiple aspects of the environment (both linguistic and extra-linguistic) contribute to language processing performance (Bahrick, Lickliter, & Flom, 2004; Kirkham, 2010; Yu & Ballard, 2007). However, theoretical and computational models of multimodal interaction during language processing have not kept pace with behavioural evidence demonstrating the complex and subtle ways in which different information sources unfold over the time-course of processing even over a single word (Ferriera & Tanenhaus, 2007; Huettig, Olivers & Hartsuiker, 2011; Anderson et al., 2011). In this paper, we provide such a model that demonstrates how information from multiple sensory modalities is integrated in the process of sentence processing. Our approach is to introduce few assumptions about the architectural constraints of such a system, but to demonstrate how constraints arise from the learner engaging with the task of mapping spoken language to the environment around her. The resulting model, thus, is compatible with unitary views of cognitive processing (Anderson, 1990), and previous modelling frameworks to investigate interactive activation of multimodal information in computation (Rogers et al., 2004), but also views that gradual development of a cognitive system in interaction with the environment is key to understanding the functioning of the mature cognitive system (Sirois et al., 2008; Westermann & Ruh, 2012).

Language mediated visual attention (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995) provides a probe to investigate how the interplay of multiple information sources from the environment affect online language processing (see Huettig, Rommers & Meyer, 2011, for review), resulting in insight to the structure of stored semantic knowledge (e.g. Huettig & Altmann, 2005; Huettig, Quinlan, McDonald & Altmann, 2006; Mirman & Magnuson, 2009; Yee & Sedivy, 2006; Yee, Overton & Thompson-Schill, 2009), stored visual knowledge (Dahan & Tanenhaus, 2005; Huettig & Altmann, 2007, 2011), stored phonological knowledge (Allopenna et al., 1998), and the temporal structure of activation of information types during spoken language comprehension (e.g. Allopenna et al., 1998; Huettig & McQueen, 2007). As gaze can be recorded with relative ease across childhood as children process visual and auditory stimuli, and does not require explicit responses to language tasks, it also enables uncovering how cognitive architecture supporting language

processing may change across the course of development (e.g. Johnson & Huettig, 2011; Mani, Johnson, McQueen & Huettig, 2013; Nadig & Sedivy, 2002; Novick, Thompson-Schill, Trueswell, 2008).

However, these behavioural studies typically (and perhaps unavoidably) lack an explicit description of how the indirect measure of eye gaze is connected to processing induced by exposure to the visual and auditory stimuli presented in such studies. The goal of the current paper is to present a word level computational model of language mediated visual attention to generate observable behaviour from the language mediated visual attention paradigm, in order to test theories of how multimodal cognitive processing results in this behaviour. Critically, the model we present is relatively unconstrained in its architecture, in order to determine the extent to which observed multimodal language processing behaviour is emergent from the structure of the environment that accompanies language comprehension situations. Interrogating the model's internal processing goes beyond the scope of current experimental and brain imaging methods in order to understand the nature of processing that emerges from theoretical assumptions about information interaction.

We first review previous behavioural, theoretical, and computational studies of multimodal language processing that have employed the language mediated visual attention paradigm. We then present a model of multimodal information interaction. We begin by determining how the model learns to map between distinct modalities, which enables a test of whether an unconstrained information processing system processes information amodally, or whether it may incorporate sensory modalities into the representations. We then examine the time-course of processing in the model as it develops, to determine how multiple information sources may interact during language processing in an emergent interactive system. We then test the model for explanatory adequacy, not only against developmental data, but also adult studies of detailed interaction of information sources during spoken word comprehension. These behavioural studies have generated theoretical descriptive accounts of the extent to which information sources influence language processing in a modular or interactionist manner, and we show how the model solves the behavioural task and the computational consequences of this solution in terms of modularity or interactivity in the model. The model therefore makes concrete current issues over the extent to which modalities in language processing are processed in a unitary or a modular manner, and generates predictions to be tested in future behavioural studies of language processing in a multimodal environment.

Behaviour and Descriptive Models of Language Mediated Visual Attention

The language mediated visual attention paradigm assesses eye movements across a visual display with concurrent speech stimuli, thus the paradigm enables a test of the interaction between multiple modalities. Manipulation of speech properties (such as the ambiguity of the speech), or the relations between observable elements of the visual display (such as the semantic and/or visual similarity of objects) enable identification of particular information sources affecting cognitive processing. Special issues dedicated to research using this paradigm were published in 2007 (Journal of Memory and Language, 57(4)) and 2011 (Acta Psychologica, 137(2)), motivated by the dramatic rise in research this field had seen over the past decade.

Although the use and application of language mediated eye gaze as a tool for exploring a broad range of questions within cognitive science has continued to increase, the mechanism by which eye gaze is affected by visual and auditory stimuli is still underspecified (Anderson et al., 2011; Ferreira & Tanenhaus, 2007; Huettig, Olivers & Hartsuiker, 2011). Gaze within such studies is typically interpreted as being drawn towards the most strongly activated item, but this is not so much an explanation as a theoretical suggestion that raises questions about activation levels, and individuating items. In order to make such a question tractable we argue it is essential to describe the nature of the information activated by the visual and auditory stimuli, the architecture that supports this activation and interaction of these two input streams and definition of the contextual boundaries in which these processes operate (e.g. to what extent are observations task specific). Developing a model of language mediated visual attention thus has broad implications, as it must offer a description of the structure of representations activated by visual and auditory streams, and the architecture that supports their activation and cross-modal interaction.

Inferring processing characteristics from visual world data

Language mediated visual attention studies using the visual world paradigm have provided evidence indicating the type of information activated as individuals process linguistic information in a rich visual environment. In alphabetic literate adult populations, items in the visual display that share their initial phonemes with the spoken target word are fixated more than items in the visual display that share phonemes in the rhyme of the word (Allopenna et al, 1998; and see Huettig, Singh & Mishra, 2011 for a description of illiterate adult behaviour). Visually displayed items that share no phonological or visual similarity with the

spoken target word, but overlap in terms of shared semantic properties, are fixated at levels that correlate with semantic similarity measures (Huettig et al., 2006; Huettig & Altmann, 2005; Mirman & Magnuson, 2009; Yee et al., 2009). Further, items in the visual display that depict visual properties of the spoken word reference, but no semantic or phonological properties, also increase gaze durations (e.g. shape: Dahan & Tanenhaus, 2005; Huettig & Altmann, 2007; colour: Huettig & Altmann, 2011). These findings have led to inferences regarding the temporal activation of information types during spoken language processing (e.g. Allopenna et al., 1998; Huettig & McQueen, 2007) and also generated a number of theoretical proposals linking gaze to the underlying processing of such stimuli.

Initial theoretical models argued that language mediated eye gaze was determined by mapping only at the level of phonological representations (phonological mapping hypothesis: Tanenhaus, Magnuson, Dahan & Chambers, 2000), or only at the level of visual representations (visual mapping hypothesis: Dahan & Tanenhaus, 2005). However, to account for all of these effects (given that, for each category of competitor, overlap in other modalities is controlled) only a model that allows gaze to be influenced by mapping at visual, semantic and phonological levels of representation is required (see Huettig & Altmann, 2005; Huettig & McQueen, 2007; Huettig, Olivers & Hartsuiker, 2011; Huettig, Olivers & Mishra, 2012).

Huettig & McQueen (2007) tested the concurrent, temporally-unfolding processing of visual, phonological, and semantic information simultaneously. They presented participants with scenes containing four objects: an item that shared its phonological onset with the spoken target word (phonological onset competitor), an item that shared semantic properties with the spoken target word (semantic competitor), an item that shared visual properties with the spoken target word (visual competitor) and an item that was not related to the spoken target word in either phonological, semantic or visual dimensions (unrelated distractor). They observed that when participants were allowed to view the display approximately 1000ms prior to target word onset, participants first looked towards the phonological onset competitor, with fixation probabilities peaking at around 400-500 ms and reducing to unrelated distractor levels once disambiguating phonemes in the spoken target word had been processed. Fixation of visual and semantic competitors were fixated with probability above unrelated distractor levels from 400-500 ms and remained so for the remainder of the recorded 1000 ms post word onset. These data are later illustrated in Figures 9a and 9b.

This pattern of fixation behaviour was interpreted as revealing cascading activation of information types through visual and speech processing systems. Huettig and McQueen (2007) suggested that during the 1000 ms preview, participants activate visual, semantic and phonological properties for each item in the display. As the spoken word unfolds, information relating to this spoken target word begins to become activated via the speech recognition system, first its phonological properties are activated which overlap with the phonological properties of the phonological onset competitor activated via the object recognition system during the preview period, leading to early fixations of this item. Information then cascades from phonological levels in the speech recognition system to activate the target word's semantic and visual properties, which overlap with the properties of visual and semantic competitors activated via the object recognition system, thus driving later fixation of these items. Note that this is *not* a (strictly) serial model as processing information types does not have to be complete before processing of other information types takes place. However, fixation of the phonological competitor before disambiguation may also, or instead, be driven by activation of its semantic and visual properties as early phonemes shared by both the phonological competitor and target word are also related to the visual and semantic properties of the phonological competitor.

Huettig and McQueen conducted a second experiment to address this issue in which exposure to the visual display prior to word onset was shortened to 200ms, to prevent activation within the object recognition system activating phonological properties of the phonological competitor prior to phonological disambiguation, and thus preventing the effects of phonological overlap on fixation behaviour. Their experiment 2 showed this was the case: phonological competitors were fixated at levels equal to unrelated distractors, but visual effects still appearing, at around 500 – 700ms post word onset, and semantic effects at approximately 700 – 900ms. Huettig & McQueen therefore concluded that these results provided strong evidence in favour of a cascaded three-level discrete mapping architecture, with phonological, then visual and semantic information influencing eye gaze (see also Yee, Huffstetler & Thompson-Schill, 2011). However, such an explanation requires fine-grained parametrisation of the time-course of information processing within each modality, and thus explicit implementation within a computational model would be required to determine the explanatory adequacy, and necessity, of such theoretical descriptive accounts.

We next identify a set of assumptions about processing of individual modalities that are potentially sufficient to generate the complex interaction of visual, semantic and phonological

effects detailed in language mediated visual attention studies. These are then implemented parsimoniously in a computational model that links explicitly the processing of the visual and auditory stimuli to eye gaze behaviour in order to test the implications and explanatory scope of such assumptions.

Information types and levels of representation

Phonological stimuli are time variant, in that information carried in the auditory signal unfolds over time. Two computational models of spoken word recognition have been applied successfully to model phonological cohort and rhyme effects on language mediated eye gaze, simulating the time-course of gradual phonological streaming (TRACE: McClelland & Elman, 1986, applied to the visual world paradigm in Allopenna et al., 1998; Magnuson et al., 2003). TRACE implements explicitly the cascading of activation from auditory to the higher levels of phonological and lexical representations. The Simple Recurrent Network (SRN) used within Magnuson et al. (2003) does not implement this hierarchy explicitly. Instead the model described in Magnuson et al. (2003) *used an error based learning algorithm to acquire the statistical dependencies among higher level phonological units and acoustic features to which the model was exposed*. The model demonstrated that the phonological effects of language mediated visual attention that were also captured by TRACE were able to be simulated with fewer architectural assumptions. Further, as an emergent model, Magnuson et al.'s (2003) simulations provided a developmental perspective, successfully replicating an observed difference in sensitivity to cohort and rhyme effects displayed by adults over the course of word learning.

In contrast with phonological information, visual stimuli in the visual world paradigm are time invariant, in that during the period gaze is recorded the visual stimulus does not change. Computational and theoretical models of visual search and object recognition generally implement hierarchical organisation of visual processing. A visual stimulus leads to activation of retinotopically organised neural assemblies sensitive to the low-level perceptual features of the visual stimulus, which in turn input to a hierarchy of layers that combine these low level features to process complex and location varying visual objects (e.g. Itti & Koch, 2001; Palmeri & Gauthier, 2004; Riesenhuber & Poggio, 2002). Visual attention models often include visual input and a distinct saliency map (e.g. Heinke & Humphreys, 2003; Mavritsaki et al., 2006, 2009; Mozer, 2002; see Itti & Koch, 2001 for review), which guides stimulus saliency and can drive fixation behaviour.

The final information type that has been shown to influence gaze in visual world studies is semantic information. Items in the visual display that share semantic relationships with the spoken target word, yet no visual or phonological relationship, are fixated more than an unrelated object (Huettig et al., 2006; Huettig & Altmann, 2005; Mirman & Magnuson, 2009; Yee & Sedivy, 2006; Yee et al., 2009). This suggests that the *semantic properties of the visual objects and/or the spoken target word become activated and recruited to influence fixation behaviour when the system is presented with their visual and auditory form, respectively*. Current models of semantic representation suggest that *a large distributed network of brain regions are involved including both regions that are modality specific (e.g. involved purely in sensory, motor or visual processing) and regions that are not modality specific yet lie at the convergence of multiple processing streams* (see Binder & Desai, 2011 for review). Although, models vary in the level to which this knowledge is embodied (i.e. sustained by perceptual processing regions) (see Barsalou, 1999; Wilson, 2002; Markman & Brendl, 2005), a unifying feature (particularly in the case of the semantic properties of concrete nouns such as those used in word level language mediated visual attention studies) is *that accessing the semantic representation of an item involves the activation of a distributed network of neural assemblies each of which encodes knowledge of distinct semantic properties of the item* (see Binder & Desai, 2011; Thompson-Schill, 2003; Martin & Chao, 2001). This has the consequence that *items that are semantically related are likely to activate similar neural structures where overlapping semantic knowledge is maintained* (e.g. Thompson-Schill, 2003).

Computational models, in particular interactive activation networks, have proved successful in displaying a broad range of properties that are also known to be displayed by the systems supporting semantic knowledge in the human brain. Hub-and-spoke models of semantic processing have proved particularly successful in modelling the behaviour of patients with impairment of their conceptual knowledge due to structural damage to brain regions associated with semantic processing (e.g. Rogers et al., 2004; Plaut 2002; Dilkina et al., 2008; 2010). The hub-and-spoke framework provides an architecture that can bring together multiple modality-specific information streams into a single system while making only minimal initial assumptions about their connectivity (see Smith, Monaghan & Huettig, 2014 for further description). The behaviour displayed by such models is thus a consequence of the structure of representations and the system's experience of the co-occurrence of information and acquisition of mappings between representations during learning stages. Semantic

knowledge of an item in such networks is therefore an emergent property of learning cross modal mappings. Mirman & Magnuson (2009) applied an interactive activation model of semantic processing (Cree & McRae, 1999; O'Connor et al., 2009) to demonstrate how the graded activation of semantic neighbours captured the complex time course of fixation behaviour displayed by participants to near and far semantic neighbours in visual world experiments. Their network represented semantic knowledge in a single layer of semantic feature units. The semantic properties of a word within the network were implemented as a unique pattern of activation within this single layer, and the trained model reflected apparent hierarchical semantic structure of concepts (see also O'Connor et al., 2009). Mirman and Magnuson's (2009) model can be conceived of as a subcomponent of the hub-and-spoke framework described above.

Models of multimodal processing

We have thus identified above the properties of the different forms of information known to be involved in language mediated visual attention, and the cognitive architectures likely to support processing of such information. In order to interpret behaviour in visual world studies a model of language mediated visual attention must describe the manner in which such information interacts. Visual world studies have provided some important evidence to inform our understanding of this debate, a debate that has been running within the field of cognitive science for many decades (e.g. see Barrett & Kurzban, 2006).

One of three categories of model could define the architecture that supports the interaction of these three knowledge types. Information processing may be modular and serial, meaning that no intermediate product of processing in one modality can influence processing in another (e.g. Fodor, 1983). This seems unlikely in the case of language mediated visual attention as we know that phonological competitors are fixated at similar levels to target items during the period prior to phonological disambiguation (see Allopenna, 1998). This suggests that processing of early phonological properties of the target word activates relationships between early phonemes and visual properties of the phonological competitor, therefore visual properties are activated before processing of the full phonological description of the target word is complete.

A second and third category of models compatible with this evidence is that information either cascades between visual and auditory processing streams (as proposed by the theoretical description of Huettig & McQueen, 2007) or that information from both streams is

processed in parallel, thus, the computational framework must distinguish the point at which information processing is unified and the points at which it remains modular, and sensory-specific.

Parallel models suggest that auditory and visual processing streams activate and integrate information from multiple modalities simultaneously, and concurrently. Such an architecture aligns with the arguments of Anderson et al. (2011) in which they propose that "to the human brain vision and language are roughly equal relevant signal streams (among several others) that can mutually constrain one another continuously in real-time". Within this framework *the cognitive system recruits information embedded in both the visual or auditory signals to constrain processing immediately*, thus cognitive processes that require the interaction of multiple modalities becomes an issue of multimodal constraint satisfaction. The hub-and-spoke framework described above provides an architecture that implements this hypothesis explicitly, integrating in parallel information from multiple modality specific input streams (Mayberry et al., 2009; Kukona & Tabor, 2011). Kukona and Tabor (2011) and Mayberry et al. (2009) both reproduced the time course of gaze patterns displayed by participants when processing ambiguous sentences using models that integrated phonological, semantic and syntactic or visual, phonological and syntactic information in parallel. However, absent from these models is the interaction of information between semantic, visual and phonological dimensions beyond the word level, which is fundamental to any explanation of the word level effects observed in studies of language mediated eye gaze (Huettig & McQueen, 2007; Huettig, Mishra & Olivers, 2012) .

Linking gaze to activation

Connecting gaze to the activation of the various information types involved in language mediated visual attention studies requires further specification (Huettig, Olivers and Hartsuiker, 2011; Kukona & Tabor, 2011; Richardson & Spivey, 2000). Altmann and Kamide (2007) state that "changes in activation state change the attentional state of the cognitive system", and this in turn "will increase the probability of a saccadic eye movement towards the spatial location associated with the change in attentional state". Similarly, Tanenhaus et al. (2000) describe the connection in terms of "automated behavioural routines that link name to its referent when the referent is visually present and task relevant, then recognising its name access these routines triggering a saccadic eye movement to fixate the relevant information". Yet, if activation causes or constitutes a shift in attention then we need

to specify how the activation of visual, semantic or phonological information becomes connected to the spatial information required to direct gaze.

Fortunately, models of visual attention have offered a detailed description of how the saliency of spatial indices, a plausible information source with which to control visual attention, can be modulated by activation of low-level visual features grounded in the information provided by the visual input (e.g. Heinke & Humphreys, 2003; Itti & Koch, 2001; Mavritsaki et al., 2006; Mozer, 2002). However, in these accounts how such information is then connected to higher-level cognitive processes such as task goals or co-activation of information in other modalities is often left unspecified. One possibility is that working memory is a key component required to bind knowledge of an item's visual, semantic, and phonological properties and associations that exist between these properties stored in long term memory, with information relating to items in the current or recent perceptual environment (Huettig, Olivers & Hartsuiker, 2011; Kahneman, Treisman & Gibbs, 1992; Kukona & Tabor, 2011; Richardson & Spivey, 2000). Alternatively, a simpler model contends that a separate short term memory component may be unnecessary, instead *short term memory is viewed in terms of the current activation landscape within a single system that connects activation in long term memory to information activated by the current visual and auditory input* (e.g. Altmann & Mirkovic, 2009; Huettig, Mishra & Olivers, 2012). These theoretical descriptive models argue that information in both the visual and auditory signal leads to activation of an item's associated linguistic and visual properties in long term memory. This activation is then able to feedback critically through the same representational substrate to visual processing levels that maintain the spatial location information relating to the item encoded in the visual input, thus grounding higher-level cognitive activity in relation to oculomotor control. To date implementations of this theoretical model have not been tested on their ability to support language mediated visual attention.

Developmental constraints

Eye gaze is used widely in developmental studies (see Trueswell, 2008 for review) as it is possible to record gaze across age groups while children perform cognitive tasks, with gaze providing a rich temporal measure of online cognitive processing. Such developmental data can provide strong tests of the explanatory adequacy of models of multimodal integration in language processing, as they must describe not only performance of the adult system but also the means by which interaction develops in a maturing system.

One major debate within the developmental literature relates to the level of detail that exists within infants' phonological representations. Specifically are phonological representations underspecified at early stages of development (e.g. Charles-luce & Luce, 1990; Metsala & Walley, 1998) or do children's phonological representations contain full segmental and sub-segmental structure from very early stages of development (e.g. Swingley & Aslin, 2000, 2002; White & Morgan, 2008; Mani & Plunkett, 2010)? Evidence put forward in support of infants possessing fully specified representations shows that 18-23 month olds direct their gaze less at an image of a target word when hearing its name mispronounced by a single segment (i.e. baby-vaby: Swingley & Aslin, 2000). A similar related study showed at 19 months that gaze towards the target varied linearly in relation to the degree of mispronunciation to the level of single phonetic feature substitutions (White & Morgan, 2008). Although, this evidence demonstrates sensitivity to the presence or absence of low-level representational features, such data does not rule out the possibility that gaze is driven by coarser grain word level representations that are less strongly activated when the auditory version of the word is adjusted.

Some models of word learning have argued that explicit training on the sub-syllabic structure of words (i.e. exposure to explicit literacy training in alphabetic languages) is required for individuals to develop phonological processing at sub-syllabic levels (Zeigler & Goswami, 2005). A recent visual world study examining the effect of phonological onset overlap and semantic overlap on fixation behaviour in illiterate adults showed that, unlike literate adults, gaze towards phonological onset competitors was not tightly time locked to the period of phonological overlap in the speech signal (Huettig, Singh & Mishra, 2011). A further study conducted using a computational model of language mediated eye gaze similar to that used in this study demonstrated that the pattern of fixation displayed by illiterates was compatible with gaze being driven by coarse grain syllable or word level representations (Smith, Monaghan & Huettig, 2014), representations that would also allow illiterates to demonstrate sensitivity to irregularities at lower levels of representation. These simulations therefore add support to arguments that prior to literacy training on alphabetic languages language mediated eye gaze may be driven by representations at a coarser grain than the phoneme.

A second related developmental debate in which studies of language mediated eye gaze have proved influential involves determining what information is activated when a child is exposed to an item's visual or phonological form at different stages of development. In a visual world study, Huang and Snedeker (2009) showed that 5 year old children, like adults, looked more

towards a competitor related to the spoken word via phonological and semantic relations, than an unrelated items. For example, when hearing the spoken word 'key' children look more towards an image of a 'log' that is phonologically related to the word 'lock, which in turn is semantically related to the spoken target word they hear 'key'. By this age, hearing the target word and seeing the objects lead to activation of phonological and semantic properties shared with other items.

We also know from related studies that from 18 months of age infants look longer at a target object's visual form when its presentation is preceded by the visual form of an object whose name is phonologically related to the target (Mani & Plunkett, 2010: e.g. prime = cat, target = cup). This indicates that exposure to the prime's visual form activates the phonological properties related to the prime, which are also shared with the target. This issue is further enlightened by evidence showing that at two years infants display phono-semantic priming, such that for example they look more at the visual form of a 'cup' when hearing 'cup' when it is preceded by viewing a 'boat' that primes phonologically related items to the target (e.g., bowl). Similar priming studies have been conducted with auditory primes (Arias-Trejo & Plunkett, 2009; Altvater-Mackensen & Mani, 2013). For example, from two years of age the visual form of a target object has been shown to be fixated longer when preceded by the auditory form of a semantically related prime item (i.e. prime = sheep, target = cow) even in conditions in which the prime is mispronounced (see Altvater-Mackensen & Mani, 2013). Together these studies provide evidence that from very early stages in development visual information in the object recognition system is progressing to activate properties of objects in phonological and semantic dimensions, and hearing the phonological form of a word leads to activation of related semantic and potentially visual properties.

As gaze can be measured across development this enables an answer to the question of whether the relative influence of overlapping representations in different modalities on language mediated eye gaze is stable over the course of development. Mani et al., (2013) demonstrated that, like adults, children as young as 2 years fixated items in the visual display that shared semantic properties with the spoken target word prior to fixating items that overlapped in terms of the target word's typical colour. This evidence demonstrates a level of stability in processing over the course of development.

The roles played by information activated via visual or auditory information have been argued to differ over the course of development. Robinson and Sloutsky (2004) suggested

that at around four years of age children display a shift in their preference from the auditory properties of stimuli to the visual properties, a bias that then continues into adulthood. Priming studies have shown that although phonological priming effects are relatively stable across development (Brooks & MacWhinney, 2000), semantic priming effects display far greater variation in younger children (Girbau & Schwartz, 2011; Plaut & Booth, 2000). Such findings motivated a recent visual world study by Mani and Huettig (submitted) that examined how sensitivity to phonological onset and semantic competitors varied over the course of development. They recorded gaze in children of 2, 4, 6 and 8 years of age as they viewed scenes containing a phonological onset competitor, a semantic competitor and two unrelated items while listening to sentences containing a spoken target word. They observed that semantic competitor effects increased in their magnitude and emerged at earlier points in the test trial as children matured. Phonological effects however displayed a more varied pattern of development, with an initial small early (at approximately 500ms post-word onset) inhibitory effect of phonological overlap on gaze at 2 years. The phonological effect became facilitatory, later post-word, and longer in duration at 4 to 6 years. By 8 years of age the effect resembled that observed previously in alphabetic literate adults, displaying increased fixation of phonological competitors in periods that mirrored the period of phonological overlap in the speech signal, with fixation of such items returning to base line levels once the speech signal disambiguated between target and competitor. The authors argue that this variation in effects over the course of development exposed an emerging prioritisation of semantic information over phonological information.

Extant models of language mediated eye gaze (Spivey, 2008; Altmann & Mirkovic, 2009; Mayberry et al., 2009; Anderson et al., 2011; Kukona & Tabor, 2011; Huettig, Mishra & Olivers, 2012) are broadly consistent with the assumption that language mediated eye gaze is a function of the interaction of knowledge stored in long term memory and the architectural constraints that define its activation via concurrent visual and auditory input streams. Changes over the course of development therefore will either result from changes to the structure of the architecture defining access to or storing information in long term memory, or changes to the structure of knowledge stored in long term memory, or both. Mayor and Plunkett (2014) demonstrate how changing the interactions between representations, but not the architecture, of TRACE was able to generate patterns of activation in its word layer units compatible with those observed in developmental studies of language mediated visual attention. They manipulated parameters within the model which represent changes to

processing within the system over the course of development, thus offering explanation for the effects observed with a detailed description of the mechanisms driving the effects. For example, the graded sensitivity to mispronunciations displayed by 19 month olds in White and Morgan (2008) could be captured by TRACE when inhibition between lexical units or phonemic units was substantially reduced. However, as TRACE is not a developmental model such parameter adjustments are reliant on additional assumptions to connect them to changes over the course of development.

Magnuson et al. (2003) modelled visual world data using a connectionist network that learned to map between phonetic features and lexical units. By charting the model's behaviour over the course of training, they demonstrated how it is able to replicate features of language mediated eye gaze behaviour displayed by adult populations as they learn novel words, by altering the processing between representations with experience of mappings between auditory and lexical information. The model offers an explicit description of how such behaviour can emerge over the course of development from a combination an associative learning mechanism and the statistical properties of the learning environment.

Although, few developmental models have been applied specifically to study language mediated eye gaze there are several models of word learning that describe the process in terms of learning cross modal mappings (e.g. McMurray, Horst, & Samuelson, 2012; Roy, 2005; Smith & Yu, 2008; Vouloumanos & Werker, 2009; Yu, Ballard, & Aslin, 2005). Such models are likely to overlap significantly with a developmental model of language mediated visual attention as they describe the knowledge likely to be recruited during language mediated eye gaze (cross modal associations) and model the statistical process through which this knowledge is extracted from the multimodal learning environment. Within these models, knowledge of a word develops as associations develop between sets of properties in separate modalities that the model is exposed to repeatedly and in a closely time-locked manner. Thus, to model and explain developing word knowledge it becomes critical to identify the structure of the multimodal input and the cross modal mappings the system is under pressure to perform. For example, Iordanescu et al. (2011) demonstrate that hearing the spoken word "cat" or hearing the sound of a cat mewing, both increase the speed with which participants are able to locate the visual depiction of a cat in a multi-object display. However, only hearing the spoken word "cat", not the sound of a cat mewing, increased the speed of locating the written word 'cat' within a display containing multiple written words. An explanation for such differences finds is compatible with such models that process associations between co-

occurring representations. Within the learning environment the visual properties of a cat mewing are likely to co-occur at a far greater frequency with the visual properties of a cat and thus the associations between these properties are likely to be higher than with the visual properties of the written form of the word cat, whereas it seems likely that the auditory properties of the spoken word cat are likely to co-occur frequently with both the properties that define a cat in its visual and orthographic form. This highlights how potentially complex properties of language mediated eye gaze can emerge from the statistical constraints imposed by the learning environment.

Study goals

In order to understand the complex processing dynamics of the multimodal system driving language mediated eye gaze it is necessary to construct explicit models of the processes that are assumed to be involved in order to link behaviour with theory. Within the current paper we implement these assumptions into a parsimonious computational model of multimodal integration for language processing. Critically, the model *learns* to integrate information between modalities, and so we examine how sensitivity of the model to semantic, phonological and visual overlap varies over the course of development, with the developmental profile providing an important constraint on explanatory adequacy of the model. The developmental analyses make two theoretical contributions. First, they allow us to determine how a model with integrated resources learn to map between a variety of modalities relevant to language processing, whether the model develops internal representations that are amodal, or whether representations are generated that are associated with particular input and output pairings (Amedi et al., 2005; Lambon Ralph & Patterson, 2008; McNorgan, Reid, McRae, 2011; Rogers et al., 2004). Second, the developmental analyses enable us to determine how the time-course of different information sources contribute to language processing alters as the model has more experience with the language.

We then examine whether the general assumptions implemented in the model are sufficient to produce the pattern of behaviour observed in studies that assess the joint contribution of visual, auditory, and semantic information in driving eye gaze behaviour in adults (Allopenna et al., 1998; Dahan & Tanenhaus, 2005; Huettig & McQueen, (2007), experiments 1 and 2; Mirman & Magnuson, 2009; Yee & Sedivy, 2006), and we relate the model's performance to theoretical descriptive accounts of information flow in language processing. This work enables us to move towards defining a baseline framework that makes explicit the connection

between the multimodal input to the language processing system and observed behavioural output (eye gaze) in order to test competing hypotheses in a robust and tractable manner. Finally, the implications of the results of this work for models of language mediated eye gaze and more broadly spoken word processing and visual attention are discussed.

2. The Multimodal Integration Model (MIM) of Language Mediated Visual Attention

Table 1: Presents visual world data replicated by the parallel multimodal integration model of language mediated visual attention (Smith, Monaghan & Huettig, 2013)

Study		Scene			
Authors	**Year**	**Item 1**	**Item 2**	**Item 3**	**Item 4**
Allopenna et al.	1998	Target	**Onset**	**Rhyme**	Distractor
Dahan & Tanenhaus	2005	Target	**Visual**	Distractor	Distractor
Huettig & Altmann	2007	**Visual**	Distractor	Distractor	Distractor
Yee & Sedivy	2006	Target	**Semantic**	Distractor	Distractor
Huettig & Altmann	2005	**Semantic**	Distractor	Distractor	Distractor
Mirman & Magnuson[a]	2009	Target	**Near Sem**	**Far Sem**	Distractor

The items displayed within scenes in each empirical study are listed with observed competitor effects highlighted in bold. Visual = visual competitor, Semantic = semantic competitor, Onset = phonological onset competitor, Rhyme = phonological rhyme competitor.

[a] *Study presented near and far semantic competitors on separate trials.*

[b] *Experiment 1.*

The MIM is a variant of the parallel multimodal integration model of language mediated visual attention that was first introduced in Monaghan and Nazir (2009) and Smith, Monaghan and Huettig, (2013; 2014). This is a parallel interactive activation model that models language mediated visual attention, and language processing more broadly in terms of multimodal constraint satisfaction. As a model of learning it charts development offering an explicit description of the connection between the structure of information available in the learning environment and language mediated eye gaze behaviour. Representations within the model emerge from a continuous process in which multiple forms of information interact in parallel rather than discrete stages of processing leading to a point of selection of an item

from a discrete candidate set (e.g. Pulvermuller et al., 2009). The mature model has previously demonstrated its ability to capture a range of word level language mediated eye gaze effects (see Table 1 and Smith, Monaghan & Huettig, 2013). The following section describes the architecture and representations used within the model.

Architecture

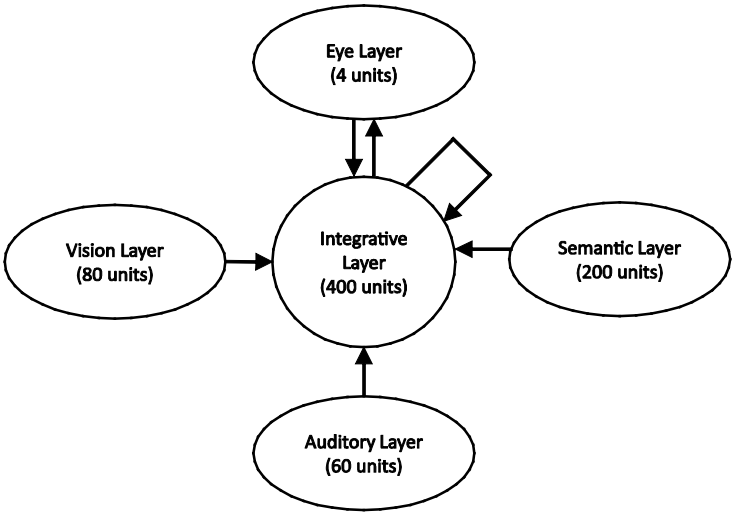

Figure 1: Architecture of the parallel multimodal integration model of language mediated visual attention (adapted from Smith, Monaghan & Huettig, 2013)

The architecture of the model was based on the parallel distributed processing framework (see Rumelhart, McClelland & the PDP Research Group, 1986; Rogers & McClelland, 2014), with the model composed of a network in which layers of non-linear processing units are connected via weighted connections (see appendix). This characteristic of the architecture is consistent with many models of phonological, visual and semantic processing (see McClelland et al., 2014, for review) and common to a number of models that have successfully captured and offer explanation for a broad range of features of language mediated visual attention (McClelland & Elman, 1986; Magnuson et al, 2003; Heinke & Humphreys, 2003; Mirman & Magnuson, 2009; Mayberry et al., 2009; Kukona & Tabor, 2011). Such a framework offers the following functional properties critical to the models success in capturing a broad range of properties of word level language mediated visual attention, it allows competition at multiple levels of representation in multiple modalities, parallel activation of representations and integration of multimodal processing streams.

The architecture of the model is presented in Figure 1. A visual processing layer consisting of 80 units allows the model to simulate visual input from up to four locations in the visual field. The layer is divided into 4 input regions of 20 units each, with regions assigned to represent the visual properties present at a given location in the visual field. The phonological layer consists of 60 units and allows the model to simulate the input of phonological information over time. The layer is divided into 6 input slots, each comprising 10 units, and each slot used to represent the unfolding phonological information input at a given point in time. Both visual and phonological layers are fully connected in a forward direction to a central integrative layer of 400 units. This central integrative layer is fully self-connected and has both full forward and back connectivity to an eye layer of 4 units and a semantic layer of 200 units. The semantic layer allows the model to simulate semantic processing and develop semantic knowledge of items. Each of the four eye layer units corresponds to one of the four locations in the visual field. The probability of fixating a given location in the visual field is therefore the level of activation of the eye layer unit corresponding to the given location. Time within the model is represented by the flow of information across weighted connections between units in the network. At each time step (ts) activation passes between all layers in the network (see appendix).

Representations

In our construction of representations we took a fundamentalist approach (Kello & Plaut, 2000: "a model should embody only the principles that are theorised to account for the phenomenon in focus"). This ensured that relationships between representations both within and across modalities were controlled. Artificial corpora consisted of 200 items with each item assigned a unique phonological, semantic and visual representation.

Visual representations were 20 unit binary feature vectors, with each unit representing the presence of absence of a given visual feature. Visual features were randomly assigned with $p = 0.5$. Semantic representations were sparsely distributed representations represented by 200 unit binary feature vectors with each unit representing a specific semantic feature. Each item was pseudo randomly assigned a unique set of 8 semantic properties.

Phonological representations were 60 unit binary feature vectors consisting of a unique sequence of 6 phonemes forming each word. Each phoneme was encoded by a 10 unit binary feature vector. A phoneme inventory of 20 phonemes was constructed for each artificial corpus, with phonological features assigned to a given phoneme randomly with p(active) =

0.5. Phonemes were then pseudo randomly sampled from the phoneme inventory to construct words of 6 phonemes in length.

Table 2: Details of relationships between targets, competitors and unrelated distractors embedded within artificial corpora. Com. = Competitor; Un. = Unrelated distractor.

Modality	Item	Constraint (Features shared with target)	Cosine Distance (μ,σ)
Phonological	Competitor	First 3 phonemes	0.260 (0.026)
	Unrelated	Max. 2 consecutive phonemes	0.500 (0.047)
Semantic	Competitor	4 of 8 semantic properties	0.500 (0.000)
	Unrelated	Max. 1 semantic property	0.961 (0.020)
Visual	Competitor	Min. 10 of 20 visual features	0.276 (0.049)
	Unrelated	Features shared with $p = 0.5$	0.518 (0.050)

We simulated the temporal component of phonological representations by presenting successive phonemes at each time step, until by word onset + 5 time steps all 60 features defining the phonological representation of the spoken word were presented as input to the phonological layer. Thus the unfolding speech signal activates a componential sequence of phonemes.

Overlap between representations was controlled as described in Table 2 ensuring that most representations had minimal overlap, but with a subset of 20 patterns having greater overlap, to reflect phonologically, semantically, or visually related representations. These 20 target items each had a phonological competitor, a semantic competitor and a visual competitor. In phonological, semantic and visual dimensions the distance between the competitor and the target was half that between the target and unrelated items. Phonological competitors shared their first three phonemes with the target word while all other phonological representations overlapped maximally by two consecutive phonemes. Finally, semantic competitors shared 4 of 8 semantic features with the target, while unrelated items shared a maximum of 1 semantic feature.

Training Procedure

Networks were trained on four tasks each lasting 14 time steps which enabled time for activation to cycle between representations in the model. Training tasks aimed to simulate tasks performed in the natural learning environment through which individuals learn cross modal mappings, mappings that are then probed in laboratory settings in visual world studies. We assume that individuals gain knowledge of an item's visual, semantic and phonological form and associations between these forms through repeated and simultaneous exposure to these multiple forms of representation (e.g., to simulate Iordanescu et al.'s (2011) data), hence training tasks produced such conditions.

Networks were trained to map from an item's visual form to its semantic form. This simulates the learning that occurs when viewing an item while simultaneously experiencing its semantic properties for example viewing a fork while eating from it. In such a training trial four items were randomly selected from the set of patterns and one was randomly assigned as the target. The visual representations of all four objects were presented as input to the visual layer at time step 0 and remained present as input for all remaining time steps. Random time invariant noise was presented as input to the phonological layer from time step 0 until the end of the training trial. The eye layer unit corresponding to the location in the visual field of the target's visual representation was fully activated also from time step 0 until the end of the test trial. Activation was free to cycle between layers until time step 3 at which point the semantic representation of the target item was clamped to the semantic layer. Error was then back propagated for the remaining time steps until the end of the trial (time step 14).

Networks were also trained to map from an item's phonological form to its semantic form. These training trials simulated the learning that occurs when an individual is simultaneously exposed to the spoken form of a word while experiencing aspects of its semantic form, for example eating from a fork while hearing the word "fork". This was simulated by first randomly selecting an item from the set of representations. At time step 0 time invariant noise was presented to the visual layer and remained present until the end of the training trial. Eye layer units were all initialised with activity 0.25, ensuring no bias in fixation of a particular location at trial onset. Also, at time step 0 the first slot of the target item's phonological representation was presented to the phonological layer of the network, at each subsequent time step an additional phoneme slot was presented to the phonological layer, until by time step 5 the full phonological representation of the target was presented as input at the phonological layer. This remained as input for all subsequent time steps. Activation within

the network was free to cycle between all layers in the network. At time step 5 the semantic representation of the target item was clamped to the semantic layer, and error back propagated for the remaining time steps until the end of the trial (time step 14).

Two tasks were also used to train the model to orientate to the visual representation of an item when its corresponding phonological or semantic properties were activated. These tasks simulated simple orientation behaviour such as the ability to fixate a fork when the concept of consuming food is activate (semantically driven orientation) or when hearing the spoken word "fork" (speech driven orientation).

During speech driven orientation training tasks, four items were randomly selected from the corpus, one of which was randomly selected as the target item. At time step 0 the visual representations of all four items were presented as input to the visual layer and remained as input for the remainder of the trial. Also at time step 0 the first phoneme slot of the phonological representation of the target was presented as input to the phonological layer. At each additional time step a further phoneme slot of the target item was presented as input at the phonological layer until by time step 5 the full representation was presented as input. At time step 5 the eye layer unit corresponding to the location of the target item was fully activated, while all other eye layer units were clamped at zero activation. Error was back propagated from time step 5 until the end of the training trial (ts = 14). Activation of all units in the semantic layer was initialised at 0 then was free to alter until the end of the training trial.

Semantically-driven orientation training trials followed a procedure similar to speech driven orientation training trials. Again, four items were randomly selected from the training corpus with a single item selected randomly as a target. At trial onset (ts = 0) the visual representations corresponding to the four randomly selected items were presented as input to the visual layer and remained as input for the remainder of the training trial. Random time invariant noise was presented as input at the phonological layer from trial onset (ts = 0) until the end of the training trial (ts = 14), simulating background noise in the local auditory environment. Also coinciding with trial onset the semantic representation of the target item was presented to the semantic layer, this remained clamped to this layer for all subsequent time steps (ts 0 – 14). At time step 3 eye layer unit corresponding to the location of the target was fully activated with all other units' activation in this layer fixed at zero. From time step 3 until the end of the training trial (ts = 14) error was back propagated.

In total 8 different simulation runs of the model were trained and tested, each initiated with a different random seed. This ensured that resulting behaviour was not a consequence of accidental aspects of the representations, starting weights, or training sequences. Connection weights in networks were initialised with random weights taken from the uniform distribution [-0.1 0.1]. New artificial corpora were created for each instantiation of the model again with the processes of construction initiated with a different random seed. Recurrent backpropagation with learning rate 0.05 was used to adjust weights during training (see appendix).

All training tasks were randomly selected and interleaved but the phonological orientation task was 4 times less likely than other tasks. This was motivated by evidence to reflect that, in the natural language learning environment, items around the child are frequently unnamed (Yu & Ballard, 2007). We therefore assumed that children are overall more likely to experience orienting to an item based on its semantic properties than based on hearing its phonological form. In total 1 million training trials were performed by each simulation run.

3. Developing properties of Language Mediated Visual Attention

In this section we examine the process through which the model acquires the necessary knowledge and behavioural routines to display language mediated eye gaze behaviour. Specifically we chart the efficacy of learning the cross modal mappings required in effectively connecting information across different modalities in language processing. We examine how the model's internal representations change over the course of acquiring these cross model mappings in order to determine whether the model solves the mapping tasks by developing amodal or modular internal representations. We then track how sensitivity to pairs of visual, semantic and phonological onset competitors emerges over the course of this development, in order to provide an explanation of developmental behavioural data and to test the model for explanatory adequacy of developing behaviour, before considering the combined contribution of all three information sources simultaneously.

Developing cross modal mapping abilities

To inform our interpretation of the behavioural effects displayed by the model when it is exposed to visual world conditions over the course of development we charted during training its level of expertise in performing semantic and spoken word orientation tasks as well as learning mappings between phonological, semantic, and visual representations.

Method

Every 50,000 training trials we tested the model on each of the four training tasks. Spoken word comprehension (Phonology -> Semantics) was tested by presenting each word's phonological representation to the phonological layer and then recording the semantic layer output after activation had been allowed to cycle through the network for 14 time steps. If the semantic layer activity was closer (lower cosine distance) to the semantic representation of the word whose phonological representation had been presented to the phonological layer than any other word in the corpus then the network was judged to have correctly comprehended the spoken word.

We also examined the model's ability to recognise objects (Vision -> Semantics). For each item in the corpus we presented its visual representation as input to one of the four locations in the model's visual field, the remaining 3 input slots were presented with random patterns with each feature p(active) = 0.5. The eye layer unit corresponding to the location of the target item was fully activated, representing focussed attention on this location in the visual field. After activation had been allowed to cycle through the network for 14 time steps activation in the semantic layer was recorded. This process was repeated with the visual representation of the target presented in each of the four possible locations in the visual field. If the semantic layer activity was closer (lower cosine distance) to that of the target item's semantic representation than any other item in the training corpus then the model was judged as correctly identifying the object.

Table 3: Accuracy of fine grain and coarse grain networks averaged over eight instantiations at the end of training for each training task (Phonology to Semantics, Vision to Semantics, Vision & Semantics to Location and Vision and Phonology to Location).

Task	**Accuracy [μ, (σ)]**
Phonology -> Semantics	100% (0.000)
Vision -> Semantics	99.4% (0.005)
Phonology -> Location	99.8% (0.004)
Semantics -> Location	99.9% (0.002)

The model was also tested at each stage of development (every 50,000 training trials) on its ability to orientate to the location of an object in the visual display when either hearing the spoken word (Phonology -> Location) or when its semantic properties were active

(Semantics -> Location). Locating an object when hearing its name was tested for each item in the corpus by first presenting the visual representation of the item as input from one of the four locations in the visual field, with input to the remaining three locations being random noise with each visual feature p(active) = 0.5. The phonological representation of the target item was presented one phoneme per time step, such that by time step 5 the full phonological representation was presented as input to the phonological layer. Activation was free to cycle within the network until time step 14 at which point activity in the eye layer was recorded. If the most activated unit in the eye layer corresponded to the location of the target's visual representation then the network was judged as correctly orientating to the target item. This was repeated so that the model was tested on its ability to orientate to each item in the corpus when presented in each of the four possible locations in the visual field. A similar procedure was used to assess the model's ability to orient to an item based on the item's semantic properties, the difference being that random noise was presented to the phonological layer instead of the target item's phonological representation, and the semantic properties of the target item were fully activated across the entire test trial defining the identity of the target.

Results

All networks learnt to perform all training tasks accurately for over 99% of items within the training corpus by the end of training (see Table 3). The accuracy of the model on each training task over the course of training is presented in Figure 2.

To examine the order in which proficiency was achieved in each task over training we examined whether the number of training trials required to reach 95% accuracy differed between tasks. This was analysed using linear mixed effects models (Baayen, 2008; Jaeger, 2008). A model with task (Phonology -> Semantics, Vision -> Semantics, Phonology & Vision -> Location, Semantics & Vision -> Location: with Phonology -> Semantics mapped onto the intercept forming the baseline condition) as a fixed effect and a random effect of simulation run of the model with random intercepts was compared to a model without the fixed effect of task. A likelihood ratio test comparison showed an effect of task ($\chi^2(3) = 63.70$, $p < 0.001$). Comparisons of means in the full model (Tukey correction for multiple comparisons) demonstrated that proficiency in mapping from phonology to semantics preceded proficiency in mapping from semantics to location (β [estimated difference between tasks in number of trials required to reach 95% accuracy threshold] = 412500, $z = 10.033$, $p < 0.001$). There was no difference in the number of trials required to gain proficiency in

mapping from vision and phonology to location (Phonology -> Location), and proficiency in mapping from vision and semantics to location (Semantics -> Location) (β = -81250, z = -1.976, p = 0.197). While, proficiency in mapping from vision to semantics resulted after proficiency in mapping from vision and semantics to location (β = 200000, z = 4.865, p < 0.001).

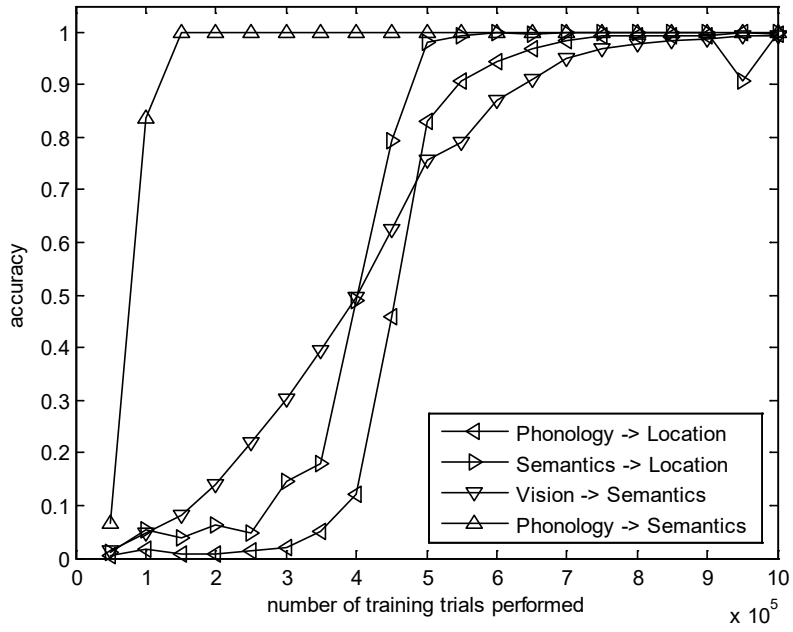

Figure 2: Accuracy of model on training tasks (Phonological -> Semantics; Visual -> Semantics; Phonological -> Location; Semantics -> Location) over the course of training.

Discussion

Crucially, our results demonstrate that the model was able to learn the various cross-modal mappings given the architectural and environmental constraints imposed in the implementation. However, these constraints lead to differences between tasks in the developmental trajectories displayed, and provide insight into the multimodal interactions that the model generates to solve the mapping tasks.

Within the model, spoken word comprehension (mapping from phonology to semantics) developed very rapidly early in development, with the model reaching proficiency in this task before any other mapping. This is the easiest cross-modal mapping task for the system to solve as it requires developing associations between information from only two of the four possible input streams and the phonological code of the target is reliable across trials. This contrasts with all other cross-modal mapping tasks which involve visual information, as these

tasks require the system to use information within either the phonological layer, semantic layer or eye layer to isolate the visual information relating to the target within the broader visual signal. For orientation tasks the system must learn to identify the visual properties of the target from its phonological or semantic properties, as well as learn the relationship between the location of this visual information and corresponding eye layer units. For visual to semantic mapping tasks the system must first learn the relationship between a given eye layer unit and the corresponding component of the visual input in order to identify the visual properties of the target that it then needs to map onto the target's corresponding semantic representation. Note that this underestimates the complexity of spoken word learning, because the model has disambiguated speech input, whereas in natural language learning situations, multiple words are likely to be spoken concurrently with viewing a scene (Estes et al., 2007; Monaghan & Mattock, 2012).

Phonological and semantic driven orientation do not differ statistically in their developmental profile within the model. Proficiency in both orientation tasks follows proficiency in spoken word comprehension and precedes proficiency in visual to semantic mapping. This may mean that the model devises similar methods for solving the orientation problem, mapping directly between the phonological and visual information, or between the semantic and visual representations, with phonology and semantic information being equally reliable for both tasks. However, as we know from the results of phonological to semantic mappings, early in development phonology activates semantics in the model, therefore the system may learn to also recruit this available information to solve the phonological orientation problem.

Vision to semantic mapping was acquired most slowly in the model. This is due to the additional complexity of this mapping task that requires the model to learn to identify the target visual representation from the four within the visual scene, and map this to the semantics. This could be resolved somewhat by implementing contingent change in input as a consequence of the model's fixation position. Very early in development, within the first six months of life, infants are able to control saccades and fixation behaviour enabling them for example to smoothly pursue a moving object (see Colombo, 2001; Trueswell, 2008 for review). When gaze changes then this changes the nature of the input to the visual system allowing infants to centre an object in its field of vision. However, the problem remains how from a range of objects the correct object is identified, or even the correct feature of the object for mapping to the word (Monaghan, Mattock, Davies, & Smith, in press; Quine, 1960).

Developing multimodal representations

To examine the representational structure that develops within networks over the course of training we investigated activity in the central integrative layer. The model may solve the task of multiple cross modal mappings by developing amodal representations in the central integrative layer that are activated when it receives information relating to a specific item irrespective of the modality of input (Rogers et al., 2004). Alternatively, the model may learn to perform the cross modal mapping without requiring such amodal representations.

Method

We first examined the extent to which the model developed location invariant representations (Stringer, Perry, Rolls, & Proske, 2006). The model receives input from four different locations in the visual field, and the model may activate a similar or a distinct pattern of units in the hidden layer for the same object presented in different locations. We calculated the cosine distance between integrative layer activity when mapping the same visual representation at each of the four visual field locations to semantics, for all pairings of visual locations. This was compared to the cosine distance in integrative layer activity between different visual representations. A positive difference in same minus different visual representation integrative layer activity indicates that hidden layer activity is more similar when processing the same item from different locations, than when processing a different item from different locations.

We also examined evidence of the development of amodal representations by comparing across trials whether activity in the central integrative layer was more similar when processing the same item but activated via different modalities (visual, auditory), compared to processing a different item input from a different modality. If amodal representations were generated by the model then integrative layer activity should be more similar when processing the same item irrespective of the modality of input. For each item in the corpus, activity in the integrative layer was recorded when mapping from phonology to semantics and when mapping from visual input to semantics for all locations in the visual field. The cosine distance between integrative layer activity when mapping from phonology to semantics and when mapping from vision to semantics was calculated for all items, for every location in the visual field. The average difference between the distance in hidden layer activity when processing different objects and hidden layer activity when processing the same object was calculated. A positive value indicates that integrative layer activity is more similar

when processing the same item (via either phonological or visual input) than hidden layer activity when processing different items (input via phonology or vision).

Both visual same minus visual different and same phonology and visual minus different phonology and visual measures were calculated at each stage of training at time step 12 (see Figure 3) of visual to semantic and/or phonology to semantic training tasks.

Results

We analysed representations at different stages of training using growth curve analysis (GCA: Singer & Willet, 2003; Magnuson, Dixon, Tanenhaus & Aslin, 2007; Mirman, Dixon & Magnuson, 2008). Growth curve analysis is a multilevel orthogonal polynomial regression technique developed to examine change over time. A level-1 model captures the overall effect of training on the difference measure plotted in Figure 3, using fourth-order orthogonal polynomials. A level-2 model then describes each level-1 model term as a function of population means, fixed effects and random effects. In this case the level-2 model describes the fixed effect of modality (phonology – visual, visual x – visual y) capturing the overall difference in the measure, and the random effect of instantiation which captures the deviation for a particular simulation for a particular modality from the mean for that particular task.

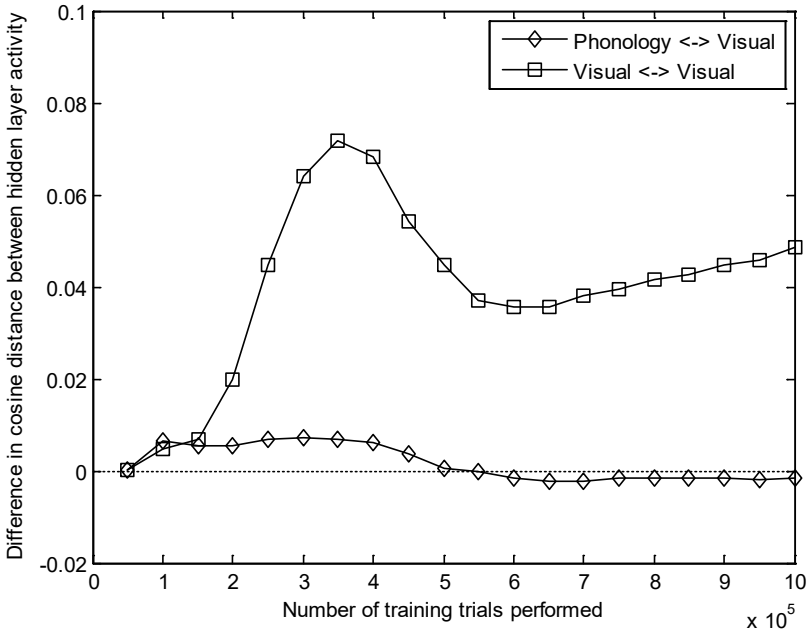

Figure 3: Difference between hidden layer activity on vis -> sem and pho -> sem task and distance between hidden layer activity on vis(location x) -> sem and vis(location y) -> sem for average item – same item examined at ts 12.

A model with a fixed effect of modality (phonology – visual, visual x – visual y: with phonology – visual mapped onto the intercept forming the baseline condition) and random effect of instantiation showed an effect of modality on the intercept ($\beta = 0.038$, $t(306) = 45.59$, $p < 0.001$), linear term ($\beta = 0.047$, $t(310) = 12.63$, $p < 0.001$), quadratic term ($\beta = -0.042$, $t(310) = -11.32$, $p < 0.001$), and cubic term ($\beta = 0.039$, $t(310) = 10.53$, $p < 0.001$). There was no effect of modality on the quartic term ($\beta = 0.003$, $t(310) = 0.89$, $p = 0.37$). The positive estimate of the intercept term indicated greater overall similarity between hidden layer activity when processing the same item with visual input from different visual locations (visual same – visual different location) than processing the same item with input from different modalities (phonology – visual). The positive estimate of the linear term indicated an increase in difference as a consequence of modality over training. Further, the significant negative effect of modality on the quadratic term indicated a steeper rise and fall in differences over training for within modality distances than between modality distances. Finally, the significant difference in the cubic term indicated a lower gradient for between modality (phonology – vision) distances around the inflection point of the development curve in comparison to the within modality (vision x – vision y) development curve.

In order to understand whether there was evidence for similarity in hidden layer processing independent of location of visual stimuli (visual x – visual y) and input modality (visual – phonology) we examined following each additional 50,000 training trials whether differences in the distance of processing the same item compared to the distance of processing a different item differed from zero (Results of these analyses can be found in appendix table A1). A positive difference indicates greater similarity in hidden layer activity when processing the same item. Results of this analysis performed on activity recorded at time step 12 of test trials shows that between 100,000 – 450,000 training trials the difference in distance between hidden layer activity when mapping from different modalities (phonology to semantics and from vision to semantics) is slightly more similar when mapping for the same item than when mapping for a different item ($t(7) > 8$, $p < 0.001$, maximum difference = 0.007 at 250,000 – 350,000 training trials). However, after 600,000 training trials this difference is slightly less similar for same items compared to different items ($t(7) > 3.4$, $p < 0.05$, maximum difference = -0.002 at 600,000 – 1,000,000 training trials). In contrasts the difference in distance between hidden layer activity when mapping from within the same modality (visual input location x to semantics and mapping from visual input location y to semantics) remains

greater when mapping for the same item than for mapping for a different item from 100,000 training trials ($t(7) > 17$, $p < 0.001$, maximum difference = 0.072 at 350,000 training trials).

Discussion

Analysis of activity within the integrative layer when processing the same visual object presented in different locations in the visual field showed that the model's representation is, from very early in training, more similar when processing the same object than processing a different object. Thus, from early stages in training the model begins to develop location invariant visual representations. There are fluctuations in the level of similarity displayed across training, with similarity peaking at around 350,000 training trials, and decreasing until approximately 600,000 training trials before similarity in processing continues to increase again for the remainder of training. Referring back to task performance over training shows that in this period (300,000-600,000) the model's proficiency on orientation tasks increases dramatically, suggesting that the model is learning to refine information regarding the location of input, but still maintaining the location invariant representation of the object in the integrative layer of the model. Location knowledge is required by the model in order to accurately direct gaze toward the location of the target.

For the model's representation of patterns from different modalities of input, the model demonstrated an initial tendency to amodal representations, where there were greater similarities than differences in the integrative layer for the same semantic representation activated from phonological or visual input. However, the model solves the task by reducing the amodal representation over time, until from 600,000 training trials onwards, the model resolves the various task constraints imposed by the learning environment by incorporating more and more aspects of the input modality into the representation of the item. This is consistent with studies of embodied cognition, where the input and output modality are intimately related to the representation itself (Barsalou, 1999; Wilson, 2002). Comparisons of activations within the integrative layer for the visual-visual and the phonological-visual representations, showed that there was greater similarity in the model's processing of the same item from the same modality than when processing the same item from different modalities. Hence, processing within the model indicates that there is greater similarity in processing an item when information regarding an item is accessed via the same modality.

Developing effects of language mediated visual attention

We next examined the point at which phonological, semantic and visual similarity effects emerge over the course of training in order to determine whether the model's emergent performance reflects that of the changes in interaction of different modalities during children's language development.

Method

Every 50,000 training trials the model was tested to examine whether the network displayed a bias towards fixating each category of competitor (phonological, semantic, visual) when each category of competitor was accompanied in the visual display by unrelated items. To test for evidence of overlap effects, activation in the eye layer was recorded when the network was exposed to the following conditions. At trial onset a scene was presented containing a competitor (phonological, visual or semantic) accompanied by random noise in the remaining three visual input slots. At time step 5 the phonological representation of the target word (related to the competitor in either a phonological, visual or semantic dimension) began to unfold. Activation in the eye layer was recorded at time steps 5, 12 and 20 providing a measure of effects at the point of word onset (ts = 5), at an early stage of word processing (ts = 12) and at a late stage of word processing (ts = 20). For all combinations of competitor and target, trials were run with the competitor presented in each of the four locations in the visual field.

Results

To examine whether competitors were fixated more than distractors we calculated the competitor bias (see equation 1) and used this as our dependent measure. Positive values indicate greater activation of the eye position associated with the competitor than eye positions associated with the visual representations of unrelated objects.

$$Competitor\ effect = \frac{activation\ of\ eye\ layer\ unit\ corresponding\ to\ competitor\ location}{\sum(activation\ of\ eye\ layer\ units\ corresponding\ to\ distractor\ locations\ /\ 3)} \quad (1)$$

Competitor effects for each type of competing stimulus (phonological bias, semantic bias, visual bias) within the model at different stages of development is displayed for time step 5 in Figure 4, for time step 12 in Figure 5 and time step 20 in Figure 6.

We used GCA to examine whether modality (phonological competitor, visual competitor, semantic competitor) affected the developmental profile of the effect on fixation behaviour.

Separate analyses were conducted on data collected at each stage of processing (ts = 5, ts = 12, ts = 20).

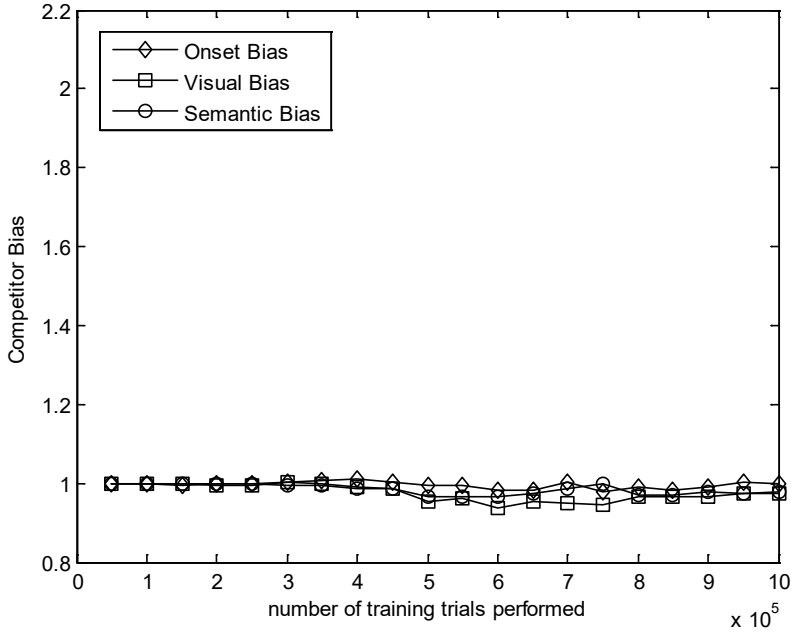

Figure 4: Magnitude of competitor (phonological [onset]; visual; semantic) effects pre-word onset (ts = 5) across training.

We tested for effects at time step 5 to examine whether any initial bias in fixation behaviour was present at word onset (see Figure 4). This was performed using a model with fixed effects of modality (semantic competitor, phonological competitor, visual competitor: with semantic competitor mapped onto the intercept forming the baseline condition) and random effects of model instantiation with random intercepts. Examining fixed effects in the model show a small yet significant difference between phonological and semantic competitors (intercept: $\beta = 0.004$, $t(21) = 4.41$, $p < 0.001$; linear: $\beta = 0.121$, $t(21) = 3.76$, $p = 0.001$; cubic: $\beta = -0.021$, $t(426) = -1.975$, $p = 0.049$). The small positive intercept estimate indicates an overall greater phonological than semantic competitor effect. The positive linear estimate indicates that the phonological competitor effect increased in relation to the semantic competitor effect over the course of training. The small yet significant negative estimate of the cubic term is likely to relate to small changes in the phonological relative to the semantic competitor effect at later stages of training. No other effects or interactions were significant ($t < 1$, $p > 0.35$). Therefore, semantic competitors did not affect behaviour of the model differently than visual competitors.

One sample t-tests for each category of competitor at each stage of training were performed. At time step 5, visual (600,000 – 1,000,000 training trials: $t(7) > 3.31$, $p < 0.02$, minimum effect = 0.903) and semantic (500,000 – 1,000,000 training trials: $t(7) > 2.43$, $p < 0.05$, minimum effect = 0.914) competitors are likely to be fixated less than unrelated distractors at later stages of training, while fixation of phonological competitors remained at the same level as distractors throughout development (see appendix Table A2 for further details).

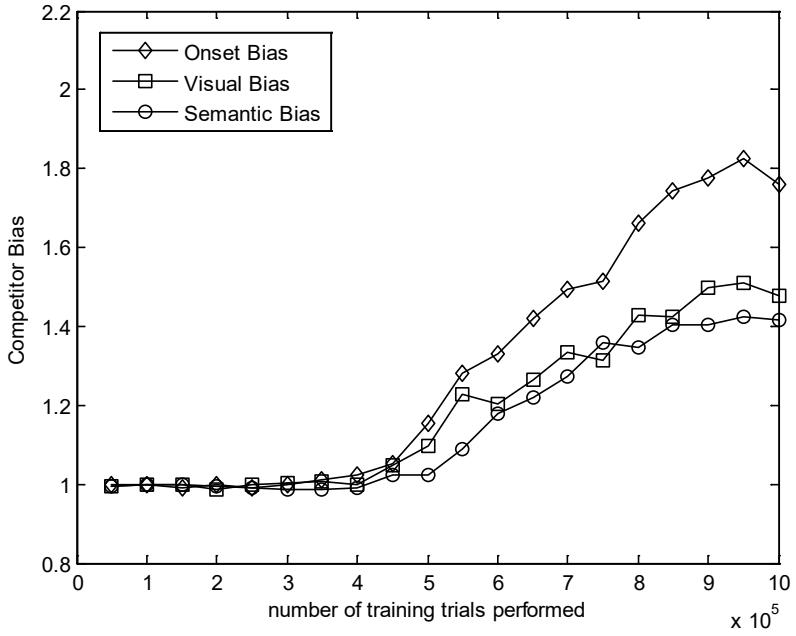

Figure 5: Magnitude of early (ts = 12) competitor (phonological [onset]; visual; semantic) effects across training.

At time step 12 (see Figure 5), a model with fixed effects of modality (semantic competitor, phonological competitor, visual competitor: with semantic competitor mapped onto the intercept forming the baseline condition) and random effects of model instantiation with random intercepts revealed an overall greater phonological than semantic competitor effect at this early stage of processing (intercept: $\beta = 0.190$, $t(23) = 4.47$, $p < 0.001$). A positive linear term indicated that the magnitude of the phonological effect relative to the semantic effect also grew at a greater rate over the course of development (linear term: $\beta = 0.732$, $t(21) = 4.172$, $p < 0.001$). The cubic term also differed significantly between phonological competitors and semantic competitors (cubic term: $\beta = -0.238$, $t(427) = -5.21$, $p < 0.001$) indicating differences towards the end of training. No other terms were significant, although the intercept and linear term capturing differences between semantic competitor and visual

competitor effects were marginal (intercept: β = 0.085, t(23) = 2.00, p = 0.058 ; linear term: β = 0.329, t(21) = 1.874, p = 0.075).

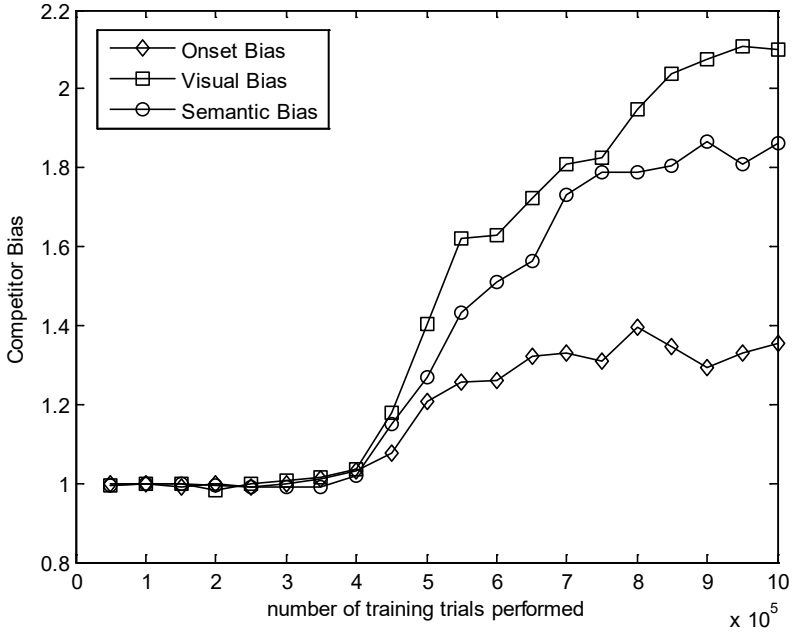

Figure 6: Magnitude of late (ts = 20) competitor (phonological [onset]; visual; semantic) effects across training.

One-sample t-tests (see appendix table A3 for full details of results), demonstrated that phonological bias was first to emerge, with the model fixating phonological competitors more than unrelated items from 400000 training trials (M = 1.045, CI = [1.013, 1.077], t(7) = 3.33, p = 0.013) and onwards (t(7) > 4.97, p < 0.003, maximum effect = 2.042). Visual competitor effects were next to emerge with visual competitors fixated more than unrelated items from 450000 trials (M = 1.110, CI = [1.008, 1.212], t(7) = 2.55, p = 0.038), onwards (t(7) > 4.74, p < 0.003, maximum effect = 1.858). Semantic competitor effects were last to emerge at this early stage in processing stimuli, with semantic competitors fixated more than distractors following 500000 training trials (M = 1.160, CI = [1.042, 1.279], t(7) = 3.207, p = 0.015) and onwards (t(7) > 4.40, p < 0.004, maximum effect = 1.697).

Finally, we conducted the same analysis on effects examined at a late processing stage (ts = 20) over the course of development (see Figure 6). The GCA results revealed a significant difference between phonological and semantic competitor effects (intercept: β = -0.375, t(23) = -6.02, p < 0.001; linear: β = -1.506, t(22) = -6.35, p < 0.001; cubic: β = 0.546, t(427) = 8.270, p < 0.001). The negative intercept estimate indicates an overall weaker phonological

than semantic competitor effect. The negative linear estimate indicates that the phonological competitor effect increased at a slower rate through development than the semantic effect. Finally, the significant positive estimate of the cubic term is likely to reflect greater changes in the phonological effect relative to the semantic effect in the tail of both functions, therefore at later stages of training. No other effects or interactions were significant ($t < 1$, $p > 0.35$). Therefore, the rate of emerging semantic competitor effects did not differ from visual competitor effects at this later stage of processing.

One sample t-tests for each modality of competitor effect (phonological, semantic, visual) at each stage of training (see appendix table A4 for full details of these results) showed that visual, semantic and phonological competitors are all first fixated more than distractors in the late time window following 400000 training trials (phonological: M = 1.049, CI = [1.016, 1.083], t(7) = 3.48, $p = 0.010$; visual: M = 1.064, CI = [1.017, 1.111], t(7) = 3.20, $p = 0.015$; semantic: M = 1.032, CI = [1.010, 1.054], t(7) = 3.49, $p = 0.010$). Visual effects ($t(7) > 4.22$, $p < 0.005$, maximum bias = 2.087) and semantic effects ($t(7) > 3.46$, $p < 0.012$, maximum bias = 2.084) remained present at all further stages of development, with visual effects being greatest at this later stage word processing. However, phonological effects peaked at 600,000 training trials (M = 1.223, CI = [1.049, 1.397], t(7) = 3.03, $p = 0.019$) and were no longer present at this late stage of word processing at later stages of training (after 650000 training trials: $t < 1.7$, $p > 0.05$).

Discussion

The above results detail the emergence of visual, semantic and phonological onset competitor effects in a parallel multimodal integration model of language mediated eye gaze that implements minimal architectural constraints on the flow of information in the system, and therefore the emergent behaviour displayed is driven by constraints imposed by the structure of information and operations imposed on the system by the learning environment. Prior to word onset, there was a small difference between semantic and phonological competitor effects, which was driven by a small negative effect of semantic similarity compared to baseline unrelated distractor levels prior to word onset, indicating that there was a slight preference for the model to orient to visual representations that were minimally overlapping in terms of meaning, prior to any other input to the model. As there was no initial positive bias in the system prior to word onset we can be confident that any later competitor effects are driven by processing of the unfolding spoken target word (phonological input).

Over the course of processing a single word, the model demonstrated that phonological similarity effects were observed prior to semantic and visual effects, when the model had processed the spoken word for only a few time steps. After longer processing time for the word, the model demonstrated greater visual and semantic effects.

However, these effects were modulated as the model had more exposure to the language, through learning trials. The model was able to reproduce developmental behavioural data that shows early in language development, phonological, semantic and visual properties of the spoken word are activated (e.g. Huang & Snedeker, 2009, Mani et al., 2012; Mani & Plunkett, 2010). Further, this description of developing effects also replicates recent visual world data showing that over the course of development semantic effects increase in magnitude and display increasingly earlier onsets, while phonological onset effects display onsets earlier in the unfolding of the spoken word relative to semantic effects and over the course of development becoming stronger and focused within earlier time windows. In addition, the model generated an untested prediction that visual similarity effects should become progressively stronger in guiding children's fixation behaviour as processing time increases, and we predict that these visual effects will emerge later in development, similar to the observed effects of semantic similarity observed later in processing a word than phonological effects in children (Mani & Huettig, submitted) .

Crucially, the complex time-course behaviour in the developmental profile of the model was entirely a consequence of a shared resource responding to the structure of information within the learning environment. As few assumptions are built into the model's architecture, it is possible to isolate the relevant properties of either the underlying system or learning environment that generate these observed effects. Thus, for instance, additional architectural constraints, such as modality-specific processing modules, or adjusting the composition or quality of the spoken language, are not required to generate these behavioural effects.

As the model integrates evidence across all modalities in parallel the item in the display whose visual properties are most strongly associated with the phonological input defining the target either directly, or indirectly via the semantics layer, will accrue activation. Initial phonemes of the target word are associated with the visual features of the phonological competitor, driving activation of the eye position associated with the phonological attractor. However, this effect is dependent on sufficient learning within the model to be able to link visual and spoken forms of words, and also to distinguish subsequences of the spoken word

as referring to particular visual representations, akin to the development of finer-grained phonological representations in the model (Smith et al., 2014). Further, we know from the analyses reported earlier in this chapter that early in development the unfolding phonological input activates semantic features consistent with the phonological input, including those of the phonological competitor. Hence, this provides a further boost to activation of the location associated with the phonologically-related visual item.

In contrast, visual and semantic effects emerge later in the processing of the word, but become increasingly early in processing the word over the course of development. This is because the visually and semantically similar items are most highly activated once all the phonological information has unfolded. This is because the phonological input will activate a semantic representation that is consistent with half the semantic features of the semantic competitor, and will prompt half the visual features of the visual competitor. Only when the entire spoken word is available is that precise link available within the model.

Thus, these complex and dynamic properties of distinct phonological, semantic and visual competitor effects do not necessarily reflect changes in the prioritisation, processing or saliency of a single category of information relative to another, instead they are likely to reflect the increasing strength of all cross-modal associations with exposure to the language, between information in visual, semantic and phonological domains.

Developing competition effects within language mediated visual attention

The above simulations assessed the competition effects from individual modalities. We next examined the effect of multiple, simultaneous competitors on performance (e.g. Huettig & McQueen, 2007; Mani et al., 2013). Such analyses enable us to test the adequacy of the model in reproducing multimodal information effects in language processing, and further, make predictions about how these combinations of information sources are integrated during development, which have not yet been fully tested in behavioural studies.

Method

To examine the interaction of competitor types across development we tested the model every 250,000 training trials (250,000; 500,000; 750,000; 1,000,000) when presented with scenes containing a visual competitor, a phonological onset competitor, a semantic competitor and an unrelated distractor, replicating the conditions of Huettig and McQueen's (2007) Experiment 1. Display onset of the visual scene was at time step 0. Onset of the target

word began at time step 5 to permit the scene to be processed prior to the word. Testing trials ran for a total of 25 time steps to permit fine-grained temporal distinctions in processing to be observed. The model was tested on each of the 20 target/competitor sets with competitors and unrelated distractors tested in each possible arrangement of location (n = 24), giving a total of 480 test trials. The model was run 8 times, with different randomisations of weights, order of training patterns, and configuration of patterns.

Results

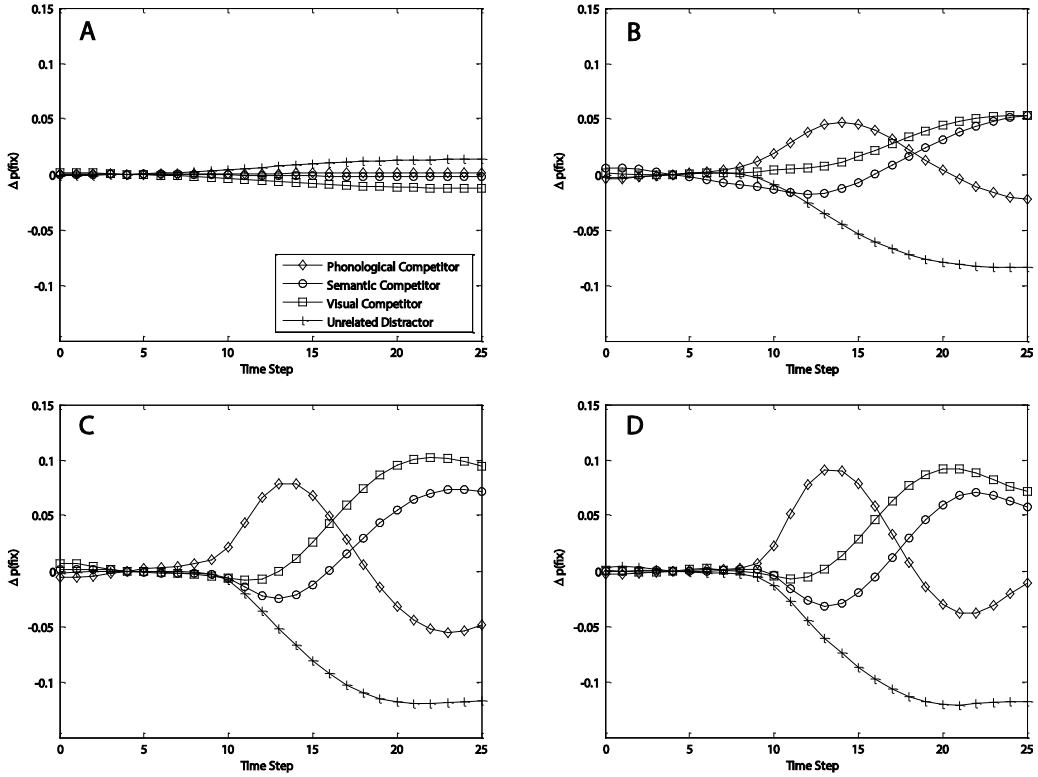

Figure 7: Proportion of fixations directed towards visual, semantic and phonological competitors after 250,000 [A]; 500,000 [B]; 750,000 [C] and 1,000,000 [D] training trials with display onset at time step 0 and word onset at time step 5.

Figure 7 presents the fixation probability of each item in the display (phonological competitor, semantic competitor, visual competitor, unrelated distractor) from word onset (ts 5) across the entire trial, averaged over all test trials (n = 480) and simulation runs (n = 8). Results are shown after 250,000, 500,000, 750,000, and 1 million training trials (Figures 10A, 10B, 10C, and 10D, respectively). The probability of fixating a given item is taken to be the level of activation of the eye unit corresponding to the location of the given item as a proportion of overall eye layer activation.

To examine the time point at which competitor effects emerge, and their magnitude at each time step across the test trial, we used one sample t-tests to examine whether the competitor bias (activation level of eye unit corresponding to location of competitor/activation level of eye unit corresponding to location of unrelated distractor) differed from 1 at each time step in the trial, for each competitor after each block of 250,000 training trials (see appendix tables A5, A6, A7, and A8 for each block of training).

This analysis revealed that after 250,000 training trials at no single time step did fixation of competitor items differ from that of unrelated distractors ($t(7) < 1.18$, $p > 0.279$). By 500,000 training trials, all competitor effects are significant. Phonological effects emerge first at time step 11 (M = 1.206, CI = [1.038, 1.373], t(7) = 2.90, p = 0.023), followed by visual effects at time step 13 (M = 1.193, CI = [1.017, 1.369], t(7) = 2.59, p = 0.036), with semantic effects last to emerge at time step 15 (M = 1.194, CI = [1.066, 1.322], t(7) = 3.58, p = 0.009). All effects remain present for all remaining time steps ($t(7) > 3.19$, $p < 0.02$). Phonological effects peak at time step 17 (M = 1.556, CI = [1.349, 1.763], t(7) = 6.07, p = 0.001), visual effects peak at time step 25 (M = 1.804, CI = [1.582, 2.026], t(7) = 8.57, p < 0.001), as do semantic effects (M = 1.771., CI = [1.505, 2.037], t(7) = 6.85, p < 0.001).

After 750,000 training trials, phonological competitor effects remain the earliest effects to be observed, by time step 10 (M = 1.150, CI = [1.030, 1.269], t(7) = 2.96, p = 0.021). Visual and semantic effects both appear from time step 13 (visual: M = 1.221, CI = [1.109, 1.334], t(7) = 4.65, p = 0.002; semantic: M = 1.128, CI = [1.014, 1.243], t(7) = 2.65, p = 0.033). Again, all effects remain present at all subsequent time steps ($t(7) > 4.74$, $p < 0.003$). Phonological effects peak at time step 16 (M = 1.933, CI = [1.699, 2.168], t(7) = 9.40, p < 0.001), while visual and semantic effects peak at time step 22 (M = 2.676, CI = [2.203, 3.148], t(7) = 8.388, p < 0.001) and time step 23 (M = 2.492, CI = [1.999, 2.985], t(7) = 7.155, p < 0.001) respectively.

Finally, after 1 million training trials, phonological effects appear by time step 10 (M = 1.198, CI = [1.051, 1.344], t(7) = 3.187, p = 0.015), visual effects are next to emerge at time step 11 (M = 1.117, CI = [1.002, 1.223], t(7) = 2.404, p = 0.047), with semantic effects last to emerge at time step 13 (M = 1.182, CI = [1.065, 1.299], t(7) = 3.685, p = 0.008). All effects remain present at all subsequent time steps ($t(7) > 3.72$, $p < 0.01$). Phonological effects peak at time step 16 (M = 2.121, CI = [1.835, 2.407], t(7) = 9.266, p < 0.001), while visual and semantic effects both peak at time step 21 (visual: M = 2.838, CI = [2.129, 3.577], t(7) =

5.821, p = 0.001; semantic: M = 2.618, CI = [2.147, 3.089], t(7) = 8.121, p < 0.001). Visual effects are the strongest of the competitor effects later in the trials, for 500,000, 750,000, and 1 million training trials.

Discussion

Behaviour of the model exposed to 1 million training trials, representing the mature language processing system, replicates the complex time course dynamics reported in adults in Huettig and McQueen's (2007) experiment 1. The model displayed an early peak in fixation of phonological competitors, with fixation then reducing throughout the rest of the trial, while semantic and visual competitor effects emerge at a later stage of processing, increasing in their magnitude over the course of the trial, until they eventually exceed levels of early phonological competitor effects. The model also replicates the numerical difference between visual and semantic effects, with visual effects slightly greater than semantic effects across the entire trial. Although this is not significant in Huettig and McQueen (2007), the numerical difference they record and later significant difference between semantic and visual effects reported in their experiment 2 is consistent with our results. In the next section of this paper we examine properties of the internal processing of the mature model and discuss how this informs our understanding of the factors driving each effect in both the model and human system.

The developmental pattern displayed with multi-competitor scenes is largely consistent with the model's performance when competitors are presented in isolation, and also consistent with the available developmental data (Mani & Huettig, submitted). Semantic effects are observed to increase in their magnitude over the course of development and emerge at increasingly earlier stages of processing, in both the model and the behavioural developmental data. The model's predictions regarding a similar developmental pattern of visual effects to date remain untested. The developmental trajectory displayed by phonological effects, increasing in magnitude and becoming increasingly time-locked to the period of phonological overlap in the speech signal between target and phonological competitor, also replicates developmental results (Mani & Huettig, submitted). From 4 to 6 years phonological effects are small and distributed across both early and late processing windows. However, by 8 years of age phonological effects begin to closely resemble that displayed by adults, peaking early in word processing and reducing rapidly in later processing periods.

Our results demonstrate that the complex time course dynamics of language mediated eye gaze competitor effects reported in adults in Huettig and McQueen (2007) experiment 1, and the developmental pattern reported in Mani & Huettig, (submitted) are emergent from a model of language mediated eye gaze, influenced by the parallel integration of information from visual, semantic and phonological modalities. The distinctive properties of each competitor type are consequences entirely of the stimulus types and learned mappings between these stimuli, and are demonstrated by the way in which the model acquires the tasks. Thus, explanations of these effects do not require changes to the architecture or structure of the learning environment over development.

4. The "Tug of War" within the mature Multimodal Integration Model (MIM) of language mediated visual attention

The above sections have demonstrated how cross modal integration becomes established within a system able to support language mediated eye gaze. We also demonstrated how the model is able to reproduce subtle time-course effects of processing multimodal stimuli (Huettig & McQueen, 2007), when provided with a preview of the visual scene prior to the spoken word commencing. However, behavioural effects are distinct when a preview of the visual scene is not available. When visual, semantic, and phonological competitor distractors are all present in the visual scene, but word onset commences at almost the same time as the scene is presented, then phonological competitor effects are no longer observed (Huettig & McQueen, 2007, experiment 2, see Figure 8b). Such effects have been interpreted in terms of modular, cascading information between modalities (Altmann & Kamide, 2007; Altmann & Mirkovic, 2009; Huettig, Olivers & Hartsuiker, 2011; Huettig, Olivers, & Mishra, 2012). When the visual objects are previewed, then this provides time for phonological information for each of these visual objects to be accessed, resulting in looks to the phonological competitor. When the visual objects are not previewed, information from the visual input does not have time to cascade to activate phonological or semantic information prior to word onset, resulting in no looks to the phonological competitor.

However, whether there is information confluence between the modalities is underspecified in descriptive accounts. That is, the time-course of activation of information – with phonological effects observed before semantic effects, which are in turn observed before visual effects – could be taken to suggest that phonology, semantics, then visual information is accessed in order (Huettig & McQueen, 2007; Yee, Huffstetler & Thompson-Schill, 2011).

Alternatively, it may be that these information sources take different times to co-influence one another, as a consequence not of architectural constraints but constraints from the representations themselves. The Multimodal Integration Model (MIM) tests this possibility.

Simulating effects of limited visual preview

In Huettig and McQueen's (2007) experiment 1, visual display onset coincided with sentence onset. On average 6.85 words were then heard by participants in the carrier sentence before the target word began to unfold. We replicated this display preview in simulations reported earlier by providing visual representations at full visual acuity from trial onset, and in addition 5 further time steps before the spoken target word begins to unfold. In Huettig and McQueen (2007) experiment 2, the preview of the display is reduced to 200ms prior to target word onset (compare Figures 9a and 9b). We now provide an implementation in the MIM to test the effect of this reduced preview.

Method

Simulations were identical to those described for the fully-trained model (after completing 1 million training trials), except that we adjusted the testing procedure to reduce the effect of visual scene preview processing in two ways. First, we simulated the effect of providing full visual acuity of all objects at the same time as the word was presented to the model (both visual and phonological information commenced at time step 5). Second, we tested the effect of requiring the model to generate high acuity visual information from all four object positions by presenting full visual acuity of the objects 10 time steps after the spoken word onset. In this condition, the spoken target word begins to unfold at time step 0, the full visual representations of all items within the display are then presented at time step 10. This was in order to accommodate the simultaneous, high-acuity processing of all four visual locations by the model. Even though object identity information can be extracted quickly from visual presentations (Altmann, 2010; Henderson & Ferreira, 2004), there is still a time delay to extracting visual information from multiple locations (see, e.g., Bar, 2007), and thus we aimed to simulate the visual input processing limitations of human participants in Huettig and McQueen's (2007) study.

Results

The fixation behaviour towards scenes containing a visual competitor, a semantic competitor, a phonological competitor and an unrelated item is displayed for each manipulation of the model in Figure 8. Figure 8c displays model fixation behavior described earlier that replicates

behavior observed in Huettig & McQueen, (2007), within these simulations scenes are presented to the model at trial onset (ts = 0), full visual acuity is achieved instantly (ts = 0) and word onset occurs at ts 5. Figure 8d shows behavior of the model when both full visual acuity and word onset occur at time step 5. Figure 8e displays behavior of the model when word onset occurs at ts 0 and full visual acuity is achieved at time step 10.

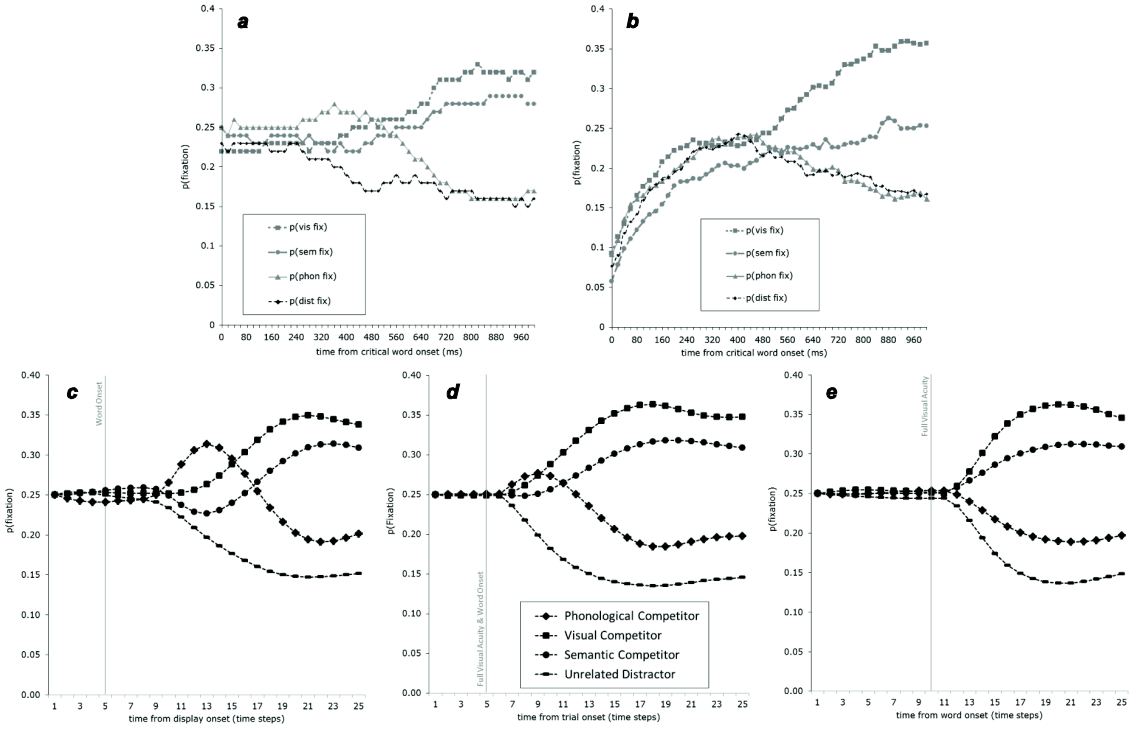

Figure 8: Probability of fixating visual, semantic and phonological competitors. a) Huettig & McQueen (2007), experiment 1 with extended preview of the visual display; b) Huettig & McQueen (2007), experiment 2, with no preview of the visual display; c) Simulations in the parallel multimodal integration model with full visual acuity at ts 0 and word onset at ts 5; d) Simulations in the parallel multimodal integration model with full visual acuity at ts 5 and word onset at ts 5; d) Simulations in the parallel multimodal integration model with full visual acuity at ts 10 and word onset at ts 0.

To examine how each of these manipulations of the model affected the magnitude of competitor effects we applied a linear mixed effects model with fixed effects of competitor (unrelated distractor, semantic competitor, visual competitor, phonological competitor: with unrelated distractor mapped onto the intercept forming the baseline condition), time window (word onset, post word onset [first 10 time steps post word onset]: with activation at word onset mapped onto the intercept forming the baseline condition), full visual acuity (ts = 0, ts

= 5, ts = 10: with ts = 0 mapped onto the intercept forming the baseline condition) and with random effects of instantiation with random intercepts and slopes. This analysis demonstrated a significant effect of time window with items fixated more post word onset than at word onset (β = 3.331, t(100) = 43.602, $p < 0.001$). Phonological (β = 1.157, t(100) = 10.71, $p < 0.001$), visual (β = 2.032, t(100) = 18.81, $p < 0.001$) and semantic (β = 1.487, t(100) = 13.76, $p < 0.001$) competitors were fixated more than unrelated items in the window post-word onset than at word onset. The difference in overall fixation of items between word onset and post-word onset windows was lower when full visual acuity occurred at time step 5 (β = -0.485, t(100) = -4.66, p < 0.001). There was a greater effect of visual (β = 1.039, t(100) = 6.80, $p < 0.001$) and semantic (β = 0.911, t(100) = 5.96, $p < 0.001$) overlap post-word onset when visual acuity occurred at time step 5, while there was no change to overall fixation of phonological competitors relative to unrelated distractors (β = -0.010, t(100) = -0.066, $p = 0.947$) for this manipulation. There was an overall greater difference in fixation of items post word onset than at word onset when full visual acuity occurred at time step 10 (β = 0.041, t(100) = 3.83, $p < 0.001$). Semantic (β = -0.038, t(100) = -2.47, $p = 0.015$), visual (β = -0.057, t(100) = -3.72, $p < 0.001$) and critically phonological effects (β = -0.071, t(100) = -4.66, $p < 0.001$) were reduced post word onset relative to word onset in the model in which full visual acuity occurred at time step 10 in comparison to the model in which full visual acuity occurred at time step 0. No other factors or interactions were significant ($t < 1.5$, $p > 0.1$).

We also examined the magnitude of competitor bias (see equation 1) at each time step and using t-tests examined whether this bias differed from 1 (see appendix table A9). A value greater than 1 indicated that the competitor was fixated above baseline unrelated distractor levels.

For the model with full visual acuity coinciding with word onset (ts = 5) phonological effects emerged at ts 3 (M = 1.118, t = 3.54, $p = 0.024$) reaching a maximum at ts 8 (M = 1.618, t = 4.348, $p = 0.012$) and returning to baseline levels between ts 13 – 19 ($t < 2.71$, $p > 0.05$). Visual effects also emerged at ts 3 (M = 1.074, t = 2.85, $p = 0.047$), remaining significant for the remainder of the trial and reaching a maximum at ts 14 (M = 2.751, t = 6.96, $p = 0.002$). Semantic effects also emerged at ts 3 (M = 1.055, t = 3.74, $p = 0.020$), remaining significant for the remainder of the test trial and reached a maximum at ts 15 (M = 2.405, t = 6.57, $p = 0.003$).

For networks that simulated full visual acuity 10 time steps after word onset, phonological effects emerged at ts 12 (M = 1.065, t = 3.54, p = 0.024), remaining significant until the end of the trial and peaked at time step 20 (M = 1.447, t = 2.84, p = 0.047). Visual effects also emerged at time step 12 (M = 1.108, t = 15.204, $p < 0.001$) and remained significant for remaining time steps, peaking at time step 20 (M = 2.738, t = 6.60, p = 0.003). Semantic effects similarly emerged at time step 12 (M = 1.101, t = 6.15, p = 0.004), remained significant for all subsequent time steps and peaked at time step 20 (M = 2.350, t = 5.993, p = 0.004).

Discussion

Our results demonstrate that the phonological effect displayed by the model can be reduced if the reduction in preview leads to less time processing a high acuity representation of the visual information in the visual display relative to word onset. In the condition in which high acuity visual information is available only from word onset there is no change in the overall magnitude of the phonological effect post word onset, however we do see a reduction in peak magnitude of the effect, further the period in which the effect exceeds that of visual and semantic effects is reduced and the period in which the effect is above baseline levels is also reduced. Under these conditions the visual input at word onset is identical in simulations of experiment 1 and 2 however it is the level of pre-processing of this visual information prior to word onset that differs between simulations and thus leads to changes in the phonological effect. The model's performance is thus consistent with theoretical proposals that when time is available for visual information to cycle through the system it pre-activates information such that information associated with early phonemes is able to exert a greater influence at early stages of word processing.

In the second set of simulations of experiment 2, in which we delay the point at which high acuity visual information pertaining to all objects in the display is achieved 10 time steps post word onset, we observe a greatly reduced overall level of phonological effect post word onset, such that at no point do phonological effects exceed that of visual or semantic effects. Such a manipulation ensures that within the model by the time visual information can influence gaze, information in the speech signal has disambiguated between target and distractor. Therefore, at no point are cross-modal associations between the concurrent visual and phonological input greater for the phonological competitor than visual or semantic competitor. The later phonemes of the word have exerted an effect to inhibit the effect of

overlap of the word's onset with the phonological distractor, but without inhibiting the influence of the visual properties of the visual competitor, or, via semantics, with the visual properties of the semantic competitor.

In the current model, the eye layer unit associated with the phonological competitor was activated slightly above distractor levels, whereas in Experiment 2 of Huettig & McQueen, (2007) there was no evidence of this. One means of aligning the model with such data is to assume that activity of a given unit in the eye layer must exceed a given threshold relative to overall eye layer activity before a saccade to any associated location is initiated, which would align the model's performance with human behaviour.

Information flow driving fixation behavior in the MIM

Huettig and McQueen (2007) interpreted the pattern of fixation behaviour with and without preview of scenes as reflecting a cascading of information through the speech recognition system. First, information from the visual display is extracted during the preview period, such that by the time the spoken word begins to unfold the visual, semantic and phonological properties of each item in the display are activated. Information from the speech stream then activates phonological properties of the spoken target word some of which are shared with that of the phonological competitor leading to early looks towards the phonological competitor. This activation then cascades through the speech recognition system to activate semantic and visual properties associated with the target word, some of which are shared with the semantic and visual competitors, hence fixation of such items increases at later stages of the trial.

In order to understand what processes drive the same pattern of fixation behaviour displayed by the model implemented in this study we examined the relationship between eye gaze and the activation of semantic information in the semantic layer of the network. In such a parallel multimodal integration model we can then address what effect does the combined visual and auditory signal have on activation in the semantic layer as the spoken target word unfolds and whether this semantic activation is then reflected in eye gaze behaviour. It is then possible to infer what information is driving activity in the eye layer within this model, and by extension predict what processes may be driving language mediated eye gaze under these conditions in visual world studies.

Method

To examine the role of activated semantic information in driving fixation behaviour within the mature model (after completing 1 million training trials) we recorded semantic layer activity at each time step of trials under conditions where the visual display contained a visual competitor, semantic competitor, a phonological onset competitor and an unrelated distractor. The visual display was provided with a preview before onset of the spoken word, as in earlier simulations of Huettig & McQueen, (2007), experiment 1. The cosine-distance between semantic layer activity and the semantic representation corresponding to the target word, the semantic competitor, the phonological competitor, the visual competitor and the unrelated distractor was then calculated for each time step. This thereby offers a measure of the level of activation of the semantic properties of each of these items over the course of word processing. The change from word onset in these distances is plotted in Figure 9 overlaid with the eye layer activity in the mature model (exposed to 1 million training trials) reported in section 3 that captured multimodal competitor effects on fixation behaviour.

Results

Using GCA we examined whether the distance between semantic layer activity and the semantic representations of competitors and targets differed from base line distractor levels post word onset (time steps 5 – 25). This was performed using a model with fixed effects of item type (unrelated distractor, target word, phonological competitor, visual competitor, semantic competitor: with unrelated distractor mapped onto the intercept forming the baseline condition) and random effects of simulation run with random intercepts. This analysis revealed that the distance between semantic layer activity and target semantic representations (intercept: $\beta = -0.656$, $t(73) = -109.66$, $p < 0.001$; linear: $\beta = -1.293$, $t(75) = -52.13$, $p < 0.001$; quadratic: $\beta = 0.958$, $t(373) = 68.81$, $p < 0.001$; cubic: $\beta = -0.189$, $t(1913) = -14.11$, $p < 0.001$; quartic: $\beta = -0.241$, $t(1913) = -17.96$, $p < 0.001$), phonological competitors' semantic representations (intercept: $\beta = -0.024$, $t(73) = -4.08$, $p < 0.001$; linear: $\beta = 0.078$, $t(75) = 3.16$, $p = 0.002$; quadratic: $\beta = 0.050$, $t(373) = 3.56$, $p < 0.001$; cubic: $\beta = -0.124$, $t(1913) = -9.25$, $p < 0.001$; quartic: $\beta = 0.079$, $t(1913) = 5.89$, $p < 0.001$) and semantic competitors' semantic representations (intercept: $\beta = -0.332$, $t(73) = -55.51$, $p < 0.001$; linear: $\beta = -0.627$, $t(75) = -25.27$, $p < 0.001$; quadratic: $\beta = 0.471$, $t(373) = 33.86$, $p < 0.001$; cubic: $\beta = -0.112$, $t(1913) = -8.36$, $p < 0.001$; quartic: $\beta = -0.105$, $t(1913) = -7.82$, $p < 0.001$) all differed from baseline unrelated distractor levels. However, distances between semantic layer activation and visual competitors' semantic representations remained at baseline unrelated distractor levels ($t < 1$,

p > 0.34). These results demonstrate that the semantic properties of the target, semantic competitor and phonological competitor were all activated above baseline unrelated distractor levels. Estimates of intercept terms show that the target's semantic properties were most strongly activated, followed by the semantic competitors, while those of the phonological competitor were slightly above baseline levels. Estimates of the linear terms indicate that the target's semantic properties were also activated at the greatest rate followed by those of the semantic competitor. The positive term on the linear estimate relating to activation of the phonological competitor's semantic properties reflects an overall decreasing activation of these properties over the course of the trial.

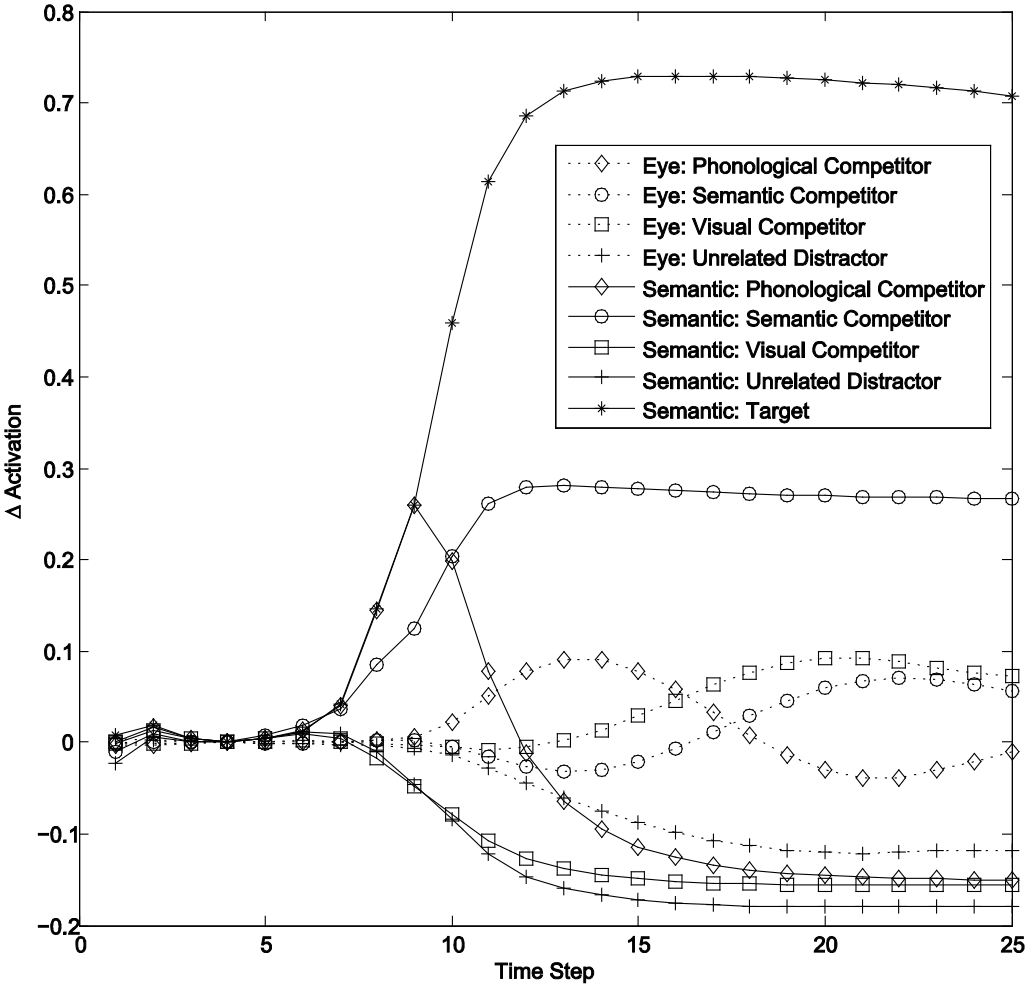

Figure 9: Eye layer and semantic layer activity during simulation of Huettig & McQueen (2007), experiment 1 in the parallel multimodal integration model trained on fine grain phonological representations

In order to determine at what points in the trial an item's semantic representation was activated above unrelated distractor levels we compared the ratio (see equation 2) between competitor (or target) distance and unrelated item distance to 1 (see appendix table A10). A ratio below 1 indicates that the competitor's (or target's) semantic representation is closer to activation in the hidden layer than that of the unrelated distractor's semantic representation.

$$Distance\ ratio\ = \frac{distance\ between\ semantic\ layer\ activity\ and\ semantic\ representation\ of\ competitor\ or\ target}{distance\ between\ semantic\ layer\ activity\ and\ semantic\ representation\ of unrelated\ distractor} \quad (2)$$

This analysis shows that the semantic properties of the semantic competitor and target item are activated above unrelated distractor levels from time step 8 ($M < 1$, $t > 2.5$, $p < 0.01$). Activation of the semantic properties of the target item peak at time steps 16 – 18 ($M = 0.083$, $t > 138$, $p < 0.001$), while activation of the semantic properties of the semantic competitor peak at time step 14 – 16 ($M = 0.542$, $t > 68.4$, $p < 0.001$). Semantic properties of the phonological competitor by contrast are activated below unrelated distractor levels prior to word onset (time step 1 – 6: $M > 1$, $t > 2.3$, $p < 0.05$) reaching a minimum at time step 4 – 5 ($M = 1.035$, $t > 2.99$, $p < 0.03$), they then increase in activation above unrelated distractor levels from time step 8 until time step 15 ($M < 1$, $t(7) < 3$, $p > 0.02$), reaching a peak of activation at time step 9 ($M = 0.666$, CI = [0.601, 0.730], $t = 12.22$, $p < 0.001$). Then from time step 16 – 25 semantic properties of the phonological competitor are activated at levels equal to those of the unrelated distractor ($t < 1.64$, $p > 0.145$). Finally, semantic properties of the visual competitor are never activated above unrelated distractor levels ($t(7) < 1$, $p > 0.5$), although from time step 2 – 9 they are activated at levels below that of the unrelated distractor ($M > 1$, $t(7) > 2.8$, $p > 0.03$), reaching a minimum level of activation at time step 8 ($M = 1.037$, CI = [1.014, 1.060], $t = 3.825$, $p = 0.007$).

Discussion

In the model, the input of the first few phonemes rapidly activates the semantic representation of both the target word and phonological distractor within the model, with activation increasing at a similar rate early in the trial for both these competitors. Activation of the semantic properties of the semantic competitor increased less rapidly, likely reflecting the fact that only half of the semantic competitor's semantic properties are associated with the unfolding phonology. When later phonemes are presented to distinguish between the phonological competitor and the target, semantic activation of the phonological competitor's

representation decreases following a peak at time step 9. By contrast, activation of the semantic properties of the target and semantic competitor continue to increase reaching asymptote following word offset at approximately time step 16. The cosine distance between semantic layer activity and the semantic representation of the target is approximately half that of the corresponding distance for the semantic competitor relative to baseline unrelated distractor levels. This suggests that as half the semantic properties of the target are shared with those of the semantic competitor it is activation of these shared properties that drives this difference in comparative activation. It is clear from the time course of eye layer activity and semantic layer activity that any activation of semantic information has a delayed effect on gaze behaviour. This is likely due to the fact that information in the semantic layer needs time to percolate down into the integrative layer, interact with information in this layer before percolating up to influence activation in the eye layer.

By contrast, at no point in the trial do the semantic properties of the visual competitor increase above baseline unrelated distractor levels, therefore the semantic properties of the visual competitor are unlikely to play a role in driving fixation of visual competitors. This suggests that in addition to indirect associations from phonology via semantics to vision there is also an influence of direct associations between phonological properties and visual properties within the model. As visual and phonological information interact directly, not mediated by semantics, such associations are likely to exert an influence on behaviour earlier than indirect associations via semantics that require information to percolate to the semantic layer and then back to the integrative layer to interact with visual information. Thus, visual effects which potentially benefit from such direct associations between phonological and visual properties should, as observed in the model, display earlier onsets than semantic effects.

Such a role for both direct and indirect associations (via semantics) influencing fixations in such a model finds further support in recent simulations that recorded semantic and eye layer activity when the model used within this paper is exposed to identical conditions but processes coarse grain phonological representations (Smith, Monaghan & Huettig, submitted). Analysis of the semantic layer in the coarse grain model shows that unlike in a model processing fine grain phonological representations activation of the semantic properties of the phonological competitor remained at baseline unrelated distractor levels at all stages of spoken word processing. This is because in the coarse grain model early properties of the speech signal are less likely to activate phonological units common to both

the phonological competitor and target word, which in the fine grain model drive activation of the semantic properties of the phonological competitor. As a result such coarse grain networks no longer display the early peak in fixation of phonological competitors. However, as seen in illiterate adults, coarse grain networks still display a weak phonological effect that is broadly distributed across the test trial. This suggests that it is the weakly activated associations between coarser grain phonological representations and the visual properties of the phonological competitor that drive these weak elongated effects in the coarse grain model.

Although a parallel interactive activation architecture is implemented in the model allowing eye gaze to be influenced by the parallel integration of information at all levels of representation and across all modalities, gaze as the earlier analysis describes is still influenced by temporal properties of the flow of information through the system and the interaction of information across modalities. For example time is required for information from visual or phonological input streams to spread through the system to activate related semantic information, and such activation percolates back through the system to interact with activated visual or phonological properties and influence fixation behaviour.

5. Summary & Conclusions

We have presented a computational model of multimodal language processing, that is able to simulate the complex interactive effects of multiple sources of information converging to affect processing as assessed in language mediated visual processing tasks. Our aim was to construct a model with few architectural assumptions, where all information sources could interact in parallel, and then to determine the extent to which the model would have to be constrained in terms of information flow to simulate observed behavioural effects. However, no such constraints were necessary. The model was able to simulate a broad range of behaviours, and in particular the precise time-course of influence of phonological, semantic, and visual information affecting an individual's language processing.

The model was trained to learn to associate pairs of representations in different modalities for the same core set of concepts. This enabled us to advance over previous work using such integrative layers to combine information from multiple sources to investigate how the model learns to assign resources to these mappings. Previous models have postulated that such an integrative resource may function as an amodal representation (Dilkina et al., 2008; 2010;

Plaut, 2002; Rogers et al., 2004) interlinking multiple modalities that are key to the representations. This was indeed the case in the early stages of the model's performance, where the model's hidden layer resulted in an activation state that was similar for different input modalities of the same item, or concept. However, as the model was provided with more exposure, it learned to distinguish its internal representations according to the modalities that were being associated, and so the input modality influenced to an increasing degree the model's processing. Though changing the resources available to the model to form mappings may affect the extent to which amodal or modally-influenced representations within the model are formed, note that the current architecture was sufficient for the same visual input but presented at different locations resulted in a high degree of similarity of activations within the model's integrative layer, and was thus not affected in the same way as representations of the same item in different modalities.

However, the central aim of the MIM was to develop a model to simulate observed multimodal effects in language processing. The model achieved this effectively, and further, was able to predict changes in the role of information sources observed in developmental studies of multimodal language processing. Emergentist modelling enables assessments of performance at multiple points of the model's history of learning, and matching the trajectory of learning to the behavioural developmental profile provides a strong test of the model's adequacy and sufficiency (Sirois et al., 2008).

The unconstraining architecture of the MIM instead describes phonological and visual effects as resulting from a combination of direct associations between phonological and visual representations and indirect associations between phonological and semantic representations, and semantic and visual representations. Semantic effects, in contrast, result from indirect associations. Thus, within the MIM no single effect can be described in terms of mapping at a single level of representation, instead they all involve the interaction of information activated across all three modalities. This parallel multimodal integration model of language mediated eye gaze thus re-frames the debate placing focus on the interaction of information across modalities (or activation of associations linking representations across modalities) rather than the activation of information within distinct modalities. Within such a framework it does not make sense to describe effects in terms of mapping at a single level of representation, as it is not possible to explain any effect within this system without describing the multimodal system as a whole. Interaction of representations at all levels is fundamental to the explanation of distinct visual, phonological or semantic effects. These results demonstrate the

hazards of inferring the properties of such a complex multimodal system without explicit implementation. As of yet no model that explains effects in terms of mapping at discrete levels has provided an explicit description of how a system is capable of generating such a pattern of effects. By contrast, the parallel multimodal integration model implemented within this paper provides an explicit description of both how eye gaze behaviour is connected to the interaction of concurrent visual and auditory stimuli with stored multimodal knowledge, and how the nature of this interaction emerges from the structure of the learning environment.

Previous theoretical models of word level effects in the visual world paradigm have framed explanations of effects in terms of mapping at discrete levels of representation, either within a single level of representation (visual: Dahan & Tanehaus, 2005; phonological: Tanenhaus, Magnuson, Dahan & Chambers, 2000) or across multiple levels (visual, semantic and phonological: Huettig & McQueen, 2007). For example, Dahan & Tanenhaus (2005) argue that "word object matching occurs at the level of visual features and not at the level of pre-activated sound forms", while Huettig & McQueen, (2007) argue that "fixations to phonological competitors are based on phonological matches thus tend to precede fixations to shape and semantic competitors". Taken together the results of the MIM demonstrate that distinct phonological, semantic and visual competitor effects observed in studies of multimodal language processing are not necessarily each generated by mappings at different discrete levels of representation. The model presented in this paper thereby fills an explanatory gap or "theoretical no man's land" (Ferriera & Tanenhaus, 2007; Huettig, Olivers & Hartsuiker, 2011; Anderson et al., 2011), describing explicitly the role of language, memory, vision and attention in connecting processing of concurrent visual and auditory signals to variation in eye gaze behaviour. Attention within the model reflects the parallel integration of evidence at all levels of representation from semantic, visual and phonological processing streams. Working memory components of the process are captured by dynamics of the recurrent activation within the model, while long term memory components are captured by weighted connections between processing units that adapt as a consequence of constraints imposed by the learning environment.

One aspect however of the model that may need addressing in future implementations relates to the visual processing component of the model. As described in section 3, this may minimally distort the developmental profile displayed by the model, specifically the connection between gaze and visual input. Mapping from vision to semantics was slow to develop in the model as visual input could not be constrained to focus on a single item.

Clearly, changes in gaze alter the nature of the visual input however, this relationship between input and gaze is not implemented in the model. We also know that infants from an early age are able to control saccades and track single objects moving in a scene. Thus it is both feasible and justifiable that to model development more accurately it may be necessary to initialize the system such that visual input varies as a function of the saliency of locations activated in the eye layer. Additional assumptions may also be required to effectively simulate the speed of object identification, as the model currently represents the visual signal in full acuity, whereas graded availability of low then high spatial frequency information may more effectively simulate relative timings of visual versus auditory processing (Bar, 2007). Extensions to the model may also address the auditory processing input to the model (Plaut & Kello, 1999). Testing the model with multi-word inputs and coarticulations across word boundaries, or adding noise to acoustic features before the phonemes themselves can be identified (see Farnetani & Recasens, 1997 for review), may affect behaviour both over the course of single word processing and over the course of development.

A further avenue of future investigation such a framework opens relates to the level to which processing impacts on perception. Within the current parsimonious implementation initial visual and phonological processing layers only feed activation forward, however it is possible to implement feedback relationships, such that processing of the input, or perceptual processing, is immediately influenced by the current multimodal activation landscape (Anderson, et al., 2011). Such an implementation would generate predictions regarding the nature of anticipatory eye gaze behavior that can then be tested in empirical studies.

With this paper we set out to establish a baseline model that identifies a finite set of assumptions common to existing models of language mediated visual that is able to generate behaviour consistent with a broad range of word level effects reported in the visual world paradigm. As a computational model it offers an explicit description of the connection between processing of concurrent visual and auditory stimuli and the distribution of eye gaze in such studies. Due to the difficulties of inferring from behaviour properties of the complex dynamic multimodal system supporting language mediated visual attention, by implementing such a model we aim to provide a baseline framework in which further assumptions of the functioning of the language processing system can be included and their implications examined.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419-439.

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*(4), 502-518.

Altmann, G., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*(4), 583-609.

Altmann, G. T. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica, 137*(2), 190-200.

Altvater-Mackensen, N., & Mani, N. (2013). The impact of mispronunciations on toddler word recognition: Evidence for cascaded activation of semantically related words from mispronunciations of familiar words. *Infancy, 18*(6), 1030-1052.

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research, 166*(3-4), 559-571.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica, 137*(2), 181-189.

Arias-Trejo, N., & Plunkett, K. (2009). Lexical–semantic priming effects during infancy. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1536), 3633-3647.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13,* 99-102.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences, 11*(7), 280-289.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: framing the debate. Psychological Review, 113(3), 628-647.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences*, 22(04), 637-660.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory.Trends in cognitive sciences, 15(11), 527-536.

Brooks, P. J., & MacWhinney, B. (2000). Phonological priming in children's picture naming. Journal of Child Language, 27(02), 335-366.

Brown-Schmidt, S., & Konopka, A. E. (2008). Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. Cognition, 109(2), 274-280.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. Journal of child Language, 17(01), 205-215.

Colombo, J. (2001). The development of visual attention in infancy. *Annual Review of Psychology*, 52, 337–367.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6(1), 84-107.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. Cognitive Science,23(3), 371-414.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. Psychonomic bulletin & review, 12(3), 453-459.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. Cognitive Neuropsychology, 25(2), 136-164.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010). Are there mental lexicons? The role of semantics in lexical decision. Brain research, 1365, 66-81.

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. Psychological Science, 18(3), 254-260.

Farnetani, E., & Recasens, D. (1997). Coarticulation and connected speech processes. The handbook of phonetic sciences, 371-404.

Ferreira, F., & Tanenhaus, M. K. (2007). Introduction to the special issue on language–vision interactions. Journal of Memory and Language, 57(4), 455-459.

Fodor, J. A. (1983). The modularity of mind: An essay on faculty psychology. MIT press.

Girbau, D., & Schwartz, R. G. (2011). Implicit Semantic priming in spanish-speaking children and adults: an auditory lexical decision task. The Spanish journal of psychology, 14(01), 4-19.

Heinke, D., & Humphreys, G. W. (2003). Attention, spatial representation, and visual neglect: simulating emergent attention and spatial memory in the selective attention for identification model (SAIM). Psychological review,110(1), 29-87.

Henderson, J. M., & Ferreira, F. (2004). Scene Perception for Psycholinguists.

Hollich, G.J., Hirsh-Pasek, K., Golinkoff, R.M., Brand, R.J., Brown, E., Chung, H.L., Hennon, E., Rocroi, C., & Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, *65*, 1-135.

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: evidence from real-time spoken language comprehension. Developmental psychology, 45(6), 1723.

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm.Cognition, 96(1), B23-B32.

Huettig, F., & Altmann, G. T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. Visual Cognition, 15(8), 985-1018.

Huettig, F., & Altmann, G. T. (2011). Looking at anything that is green when hearing “frog”: How object surface colour and stored object colour knowledge influence language-mediated overt attention. The Quarterly Journal of Experimental Psychology, 64(1), 122-145.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. Journal of Memory and Language, 57(4), 460-482.

Huettig, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. T. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. Acta psychologica, 121(1), 65-80.

Huettig, F., Mishra, R. K., & Olivers, C. N. (2012). Mechanisms and representations of language-mediated visual attention. Frontiers in psychology, 2.394.

Huettig, F., Olivers, C. N., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. Acta psychologica, 137(2), 138-150.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. Acta psychologica, 137(2), 151-171.

Huettig, F., Singh, N., & Mishra, R. K. (2011). Language-mediated visual orienting behavior in low and high literates. Frontiers in psychology, 2.

Iordanescu, L., Grabowecky, M., & Suzuki, S. (2011). Object-based auditory facilitation of visual search for pictures and words with frequent and rare targets. Acta psychologica, 137(2), 252-259.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. Nature reviews neuroscience, 2(3), 194-203.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of memory and language, 59(4), 434-446.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. Cognitive psychology, 24(2), 175-219.

Kello, C. T., & Plaut, D. C. (2000). Strategic control in word reading: evidence from speeded responding in the tempo-naming task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(3), 719-750.

Kirkham, N.Z. (2010) Altogether now: Learning through multiple sources. In S.P. Johnson (Ed.), *Neoconstructivism: The new science of cognitive development.* New York: Oxford University Press.

Kukona, A., & Tabor, W. (2011). Impulse processing: A dynamical systems model of incremental eye movements in the visual world paradigm. Cognitive science, 35(6), 1009-1051.

Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and differentiation in semantic memory. Annals of the New York Academy of Sciences, 1124(1), 61-76.

Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. Neuron, 62(2), 281-290.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: studies with artificial lexicons. Journal of Experimental Psychology: General, 132(2), 202.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. Cognitive Science, 31(1), 133-156.

Mani, N., & Huettig, F. (submitted). The changing dynamics of word-referent mapping across development.

Mani, N., & Plunkett, K. (2010). In the infant's mind's ear evidence for implicit naming in 18-month-olds. Psychological science, 21(7), 908-913.

Mani, N., Durrant, S., & Floccia, C. (2012). Activation of phonological and semantic codes in toddlers. Journal of Memory and Language, 66(4), 612-622.

Mani, N., Johnson, E., McQueen, J. M., & Huettig, F. (2013). How yellow is your banana? Toddlers' language-mediated visual search in referent-present tasks. Developmental psychology, 49(6), 1036.

Markman, A. B., & Brendl, C. M. (2005). Constraining theories of embodied cognition. *Psychological Science*, 16(1), 6-10.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: structure and processes. Current opinion in neurobiology, 11(2), 194-201.

Mavritsaki, E., Heinke, D., Humphreys, G. W., & Deco, G. (2006). A computational model of visual marking using an inter-connected network of spiking neurons: The spiking search over time & space model (sSoTS).Journal of Physiology-Paris, 100(1), 110-124.

Mavritsaki, E., Allen, H., & Humphreys, G. (2009). Model based analysis of fMRI-data: applying the sSoTS framework to the neural basic of preview search. In Attention in cognitive systems (pp. 124-138). Springer Berlin Heidelberg.

Mavritsaki, E., Allen, H. A., & Humphreys, G. W. (2010). Decomposing the neural mechanisms of visual search through model-based analysis of fMRI: top-down excitation, active ignoring and the use of saliency by the right TPJ.Neuroimage, 52(3), 934-946.

Mavritsaki, E., Heinke, D., Allen, H., Deco, G., & Humphreys, G. W. (2011). Bridging the gap between physiology and behavior: evidence from the sSoTS model of human visual attention. Psychological review, 118(1), 3.

Mayberry, M. R., Crocker, M. W., & Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. Cognitive science, 33(3), 449-496.

Mayor, J., & Plunkett, K. (2014). Infant word recognition: Insights from TRACE simulations. Journal of memory and language, 71(1), 89-123.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive psychology, 18(1), 1-86.

McClelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition.Cognitive science, 38(6), 1139-1189.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). Parallel distributed processing. Explorations in the microstructure of cognition, 2, 216-271.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831-877.

McNorgan, C., Reid, J., & McRae, K. (2011). Integrating conceptual knowledge within and across representational modalities. Cognition, 118(2), 211-233.

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. Memory & cognition, 37(7), 1026-1039.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. Journal of memory and language, 59(4), 475-494.

Monaghan, P. & Nazir, T. (2009). Modelling sensory integration and embodied cognition in a model of word recognition. In J. Mayor, N. Ruh, & K. Plunkett (Eds.), *Connectionist models of behaviour and cognition II*, pp.337-348. Singapore: World Scientific.

Moore, C., Angelopoulos, M., & Bennett, P. (1999). Word learning in the context of referential and salience cues. *Developmental Psychology, 35*(1), 60-68.

Mozer, M. C. (2002). Frames of reference in unilateral neglect and spatial attention: A computational perspective. Psychological Review, 109, 156-185.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. Psychological Science,13(4), 329-336.

Novick, J. M., Thompson-Schill, S. L., & Trueswell, J. C. (2008). Putting lexical constraints in context into the visual-world paradigm. Cognition, 107(3), 850-903.

O'Connor, C. M., Cree, G. S., & McRae, K. (2009). Conceptual hierarchies in a flat attractor network: Dynamics of learning and computations. Cognitive Science, 33(4), 665-708.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. Nature Reviews Neuroscience, 5(4), 291-303.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, 1(2), 263–269.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. Psychological review, 107(4), 786-823.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In MacWhinney, B. (Ed.), *The emergence of language*, (pp. 381-415). New York, NY: Taylor & Francis.

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: A computational account of optic aphasia. Cognitive Neuropsychology, 19(7), 603-639.

Pulvermüller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. Brain and language, 110(2), 81-94.

Quine, W. V. O. (1960). Word and object. Cambridge, MA: MIT Press.

R Development Core Team. (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition, 76*(3), 269-295.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*(2), 162-168.

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development, 75*(5), 1387-1401.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science, 38*(6), 1024-1077.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological Review, 111*(1), 205-235.

Schenk, T., & McIntosh, R. D. (2010). Do we have independent visual streams for perception and action? *Cognitive Neuroscience, 1*(1), 52-62.

Shipp, S. (2004). The brain circuitry of attention. *Trends in Cognitive Sciences, 8*(5), 223-230.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford university press.

Sirois, S., Spratling, M., Thomas, MSC., Westermann, G., Mareschal, D., & Johnson, M. (2008). Précis of Neuroconstructivism. *Behavioral and Brain Sciences, 31*, 321-331.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558-1568.

Smith, A., Monaghan, P., & Huettig, F. (2013). An amodal shared resource model of language-mediated visual attention. *Frontiers in Psychology, 4*, 528.

Smith, A., Monaghan, P., & Huettig, F. (2014). Literacy effects on language and vision: Emergent effects from an amodal shared resource (ASR) computational model. *Cognitive Psychology, 75*, 28-54.

Smith, Monaghan & Huettig, (submitted). *Complex word recognition behaviour emerges from the richness of the word learning environment*.

Spivey, M. (2008). The Continuity of Mind. New York, NY: Oxford University Press.

Stringer, S. M., Perry, G., Rolls, E. T., & Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics*, *94*(2), 128-142.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition, 76*(2), 147-166.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science, 13*(5), 480-484.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research, 29*(6), 557-580.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632-1634.

Thompson-Schill, S. L. (2003). Neuroimaging studies of semantic memory: inferring "how" from "where". *Neuropsychologia, 41*(3), 280-292.

Trueswell, J. C. (2008). Using eye movements as a developmental measure within psycholinguistics. Language acquisition and language disorders, 44, 73.

Van de Velde, M., Meyer, A. S., & Konopka, A. E. (2014). Message formulation and structural assembly: describing "easy" and "hard" events with preferred and dispreferred syntactic structures. *Journal of Memory and Language, 71*(1), 124-144.

Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology, 45*(6), 1611-1617.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(3), 387-401.

Westermann, G., & Ruh, N. (2012). A neuroconstructivist model of past tense development and processing. *Psychological Review, 119,* 649-667.

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language, 59*(1), 114-132.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4), 625-636.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 1-14.

Yee, E., Huffstetler, S., & Thompson-Schill, S. L. (2011). Function follows form: Activation of shape and function features during object identification. *Journal of Experimental Psychology: General*, 140(3), 348.

Yee, E., Overton, E., & Thompson-Schill, S. L. (2009). Looking for meaning: Eye movements are sensitive to overlapping semantic features, not association. *Psychonomic Bulletin & Review, 16*(5), 869-874.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70*(13), 2149-2165.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science, 29*(6), 961-1005.

Yurovsky, D., Boyer, T. W., Smith, L. B., & Yu, C. (2013). Probabilistic cue combination: less is more. *Developmental Science*, *16*(2), 149-158.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. Psychological Bulletin, 131(1), 3.

Appendix

Neural networks simulations were conducted using Mikenet version 8.0 developed by M. W. Harm (www.cnbc.cmu.edu/~mharm/research/tools/mikenet/), a collection of libraries written in the C programming language for implementing and training connectionist networks.

Networks were trained using the continuous recurrent backpropagation through time training algorithm provided in Mikenet (crbp.c) which implements Pearlmutter (1989). Unit activation was calculated using a logistic activation function and sum squared error was used to calculate error. Time within the network was modelled using 14 samples and an integration constant of 0.25. All other parameters were set to the default values implemented in Mikenet version 8.0.

Mixed effects model analysis was performed using the R (version 3.1.0; R Development Core Team, 2009) libraries lme4 (version 1.1-6) and languageR (version 1.4.1).

Table A1: Development of amodal representations ts 12

Fine Grain

Competitor	Training Stage	Diff.	Confidence Interval Lower	Upper	t-value	p-value
Phon <-> Vis	50000	0.000	0.000	0.000	1.033	0.336
	100000	0.006	0.006	0.007	31.111	0.000
	150000	0.006	0.005	0.007	13.389	0.000
	200000	0.005	0.005	0.006	17.745	0.000
	250000	0.007	0.006	0.008	18.845	0.000
	300000	0.007	0.006	0.008	21.405	0.000
	350000	0.007	0.006	0.008	23.228	0.000
	400000	0.006	0.005	0.007	21.170	0.000
	450000	0.004	0.003	0.004	15.626	0.000
	500000	0.001	-0.001	0.002	1.062	0.324
	550000	0.000	-0.001	0.001	-0.589	0.574
	600000	-0.002	-0.003	-0.001	-4.262	0.004
	650000	-0.002	-0.003	-0.001	-4.671	0.002
	700000	-0.002	-0.003	-0.002	-7.592	0.000
	750000	-0.002	-0.002	-0.001	-5.279	0.001
	800000	-0.002	-0.002	-0.001	-6.870	0.000
	850000	-0.002	-0.003	-0.001	-3.637	0.008
	900000	-0.002	-0.003	-0.001	-3.871	0.006
	950000	-0.002	-0.003	-0.001	-3.817	0.009
	1000000	-0.002	-0.003	-0.001	-3.415	0.011
Vis <-> Vis	50000	0.000	0.000	0.000	2.306	0.055
	100000	0.005	0.004	0.005	23.644	0.000
	150000	0.007	0.006	0.008	17.382	0.000
	200000	0.020	0.018	0.021	30.231	0.000
	250000	0.045	0.042	0.048	34.319	0.000
	300000	0.064	0.060	0.068	42.472	0.000
	350000	0.072	0.067	0.076	38.846	0.000
	400000	0.068	0.065	0.072	43.770	0.000
	450000	0.054	0.051	0.058	37.352	0.000
	500000	0.045	0.040	0.049	23.997	0.000
	550000	0.037	0.034	0.040	25.474	0.000
	600000	0.036	0.032	0.040	21.280	0.000
	650000	0.036	0.032	0.040	21.076	0.000
	700000	0.038	0.034	0.042	21.559	0.000
	750000	0.040	0.035	0.044	21.550	0.000
	800000	0.042	0.038	0.046	24.855	0.000
	850000	0.043	0.037	0.048	18.928	0.000
	900000	0.045	0.039	0.050	18.634	0.000
	950000	0.046	0.040	0.052	18.063	0.000
	1000000	0.048	0.042	0.055	18.658	0.000

Table A2: Development of pre-word onset competitor effects

Fine Grain

Competitor	Training Stage	Comp. Bias	Confidence Interval Lower	Upper	t-value	p-value
Phonological	50000	0.998	0.996	1.000	-2.823	0.026
	100000	0.997	0.993	1.000	-2.542	0.039
	150000	0.999	0.994	1.004	-0.431	0.680
	200000	1.005	0.999	1.010	2.076	0.077
	250000	1.004	0.997	1.010	1.252	0.251
	300000	1.004	0.995	1.014	1.136	0.293
	350000	1.005	0.998	1.013	1.669	0.139
	400000	1.012	0.998	1.026	1.948	0.092
	450000	1.016	0.989	1.043	1.403	0.203
	500000	1.009	0.979	1.038	0.704	0.504
	550000	1.011	0.981	1.041	0.846	0.425
	600000	1.010	0.973	1.046	0.638	0.544
	650000	0.992	0.967	1.017	-0.752	0.476
	700000	1.008	0.976	1.040	0.586	0.576
	750000	0.998	0.959	1.036	-0.140	0.893
	800000	0.997	0.947	1.048	-0.128	0.902
	850000	1.001	0.963	1.040	0.072	0.945
	900000	0.992	0.952	1.032	-0.480	0.646
	950000	0.988	0.947	1.028	-0.710	0.500
	1000000	1.000	0.967	1.034	0.005	0.997
Visual	50000	1.001	0.998	1.004	0.910	0.393
	100000	1.000	0.996	1.004	0.126	0.903
	150000	1.002	1.000	1.004	2.862	0.024
	200000	0.998	0.994	1.003	-0.917	0.390
	250000	1.002	0.997	1.008	0.954	0.372
	300000	1.001	0.995	1.007	0.444	0.670
	350000	0.998	0.992	1.003	-1.105	0.306
	400000	0.994	0.983	1.005	-1.333	0.224
	450000	0.992	0.961	1.024	-0.581	0.579
	500000	0.965	0.938	0.992	-3.085	0.018
	550000	0.966	0.915	1.017	-1.597	0.154
	600000	0.946	0.921	0.970	-5.262	0.001
	650000	0.932	0.903	0.960	-5.642	0.001
	700000	0.913	0.875	0.952	-5.280	0.001
	750000	0.933	0.887	0.979	-3.423	0.011
	800000	0.917	0.865	0.968	-3.845	0.006
	850000	0.908	0.850	0.966	-3.750	0.007
	900000	0.914	0.868	0.960	-4.466	0.003
	950000	0.915	0.855	0.976	-3.313	0.013
	1000000	0.903	0.846	0.959	-4.094	0.005
Semantic	50000	1.000	0.997	1.003	-0.246	0.813
	100000	0.999	0.996	1.003	-0.358	0.731
	150000	1.000	0.996	1.003	-0.145	0.889
	200000	0.998	0.996	0.999	-3.102	0.017
	250000	0.999	0.995	1.003	-0.780	0.461
	300000	0.998	0.995	1.001	-1.791	0.116
	350000	0.992	0.984	1.000	-2.488	0.042
	400000	0.987	0.973	1.001	-2.151	0.069
	450000	0.986	0.969	1.002	-2.042	0.081
	500000	0.966	0.944	0.987	-3.765	0.007
	550000	0.950	0.901	0.999	-2.426	0.046
	600000	0.952	0.907	0.997	-2.532	0.039
	650000	0.933	0.903	0.963	-5.231	0.001
	700000	0.930	0.881	0.978	-3.414	0.011
	750000	0.938	0.896	0.980	-3.529	0.010
	800000	0.932	0.883	0.982	-3.243	0.014
	850000	0.942	0.886	0.998	-2.464	0.043
	900000	0.920	0.873	0.968	-3.936	0.006
	950000	0.935	0.901	0.969	-4.544	0.003
	1000000	0.914	0.883	0.945	-6.503	0.000

Table A3: Development of early competitor effects

Fine Grain

Competitor	Training Stage	Comp. Bias	Confidence Interval Lower	Upper	t-value	p-value
Phonological	50000	0.995	0.991	1.000	-2.204	0.063
	100000	0.994	0.988	1.000	-2.483	0.042
	150000	0.996	0.982	1.009	-0.711	0.500
	200000	1.009	0.996	1.021	1.633	0.147
	250000	1.004	0.987	1.021	0.555	0.596
	300000	1.009	0.986	1.031	0.885	0.406
	350000	1.015	0.997	1.033	1.988	0.087
	400000	1.045	1.013	1.077	3.333	0.013
	450000	1.155	1.106	1.204	7.469	0.000
	500000	1.255	1.134	1.376	4.973	0.002
	550000	1.405	1.309	1.501	9.956	0.000
	600000	1.556	1.376	1.737	7.296	0.000
	650000	1.651	1.468	1.835	8.405	0.000
	700000	1.836	1.651	2.021	10.698	0.000
	750000	1.911	1.677	2.146	9.183	0.000
	800000	1.920	1.712	2.127	10.492	0.000
	850000	1.966	1.783	2.148	12.499	0.000
	900000	1.940	1.739	2.140	11.095	0.000
	950000	1.997	1.730	2.263	8.837	0.000
	1000000	2.042	1.829	2.254	11.594	0.000
Visual	50000	1.001	0.992	1.010	0.279	0.788
	100000	1.001	0.993	1.009	0.272	0.793
	150000	1.006	0.999	1.012	2.012	0.084
	200000	0.998	0.988	1.007	-0.600	0.568
	250000	1.013	0.998	1.028	1.997	0.086
	300000	1.008	0.987	1.029	0.891	0.403
	350000	1.014	0.997	1.031	1.912	0.098
	400000	1.034	0.993	1.075	1.979	0.088
	450000	1.110	1.008	1.212	2.548	0.038
	500000	1.231	1.116	1.346	4.742	0.002
	550000	1.336	1.225	1.447	7.143	0.000
	600000	1.375	1.260	1.490	7.685	0.000
	650000	1.473	1.272	1.674	5.568	0.001
	700000	1.455	1.265	1.645	5.670	0.001
	750000	1.626	1.469	1.783	9.407	0.000
	800000	1.714	1.552	1.877	10.410	0.000
	850000	1.742	1.646	1.839	18.214	0.000
	900000	1.765	1.635	1.896	13.859	0.000
	950000	1.858	1.688	2.028	11.929	0.000
	1000000	1.841	1.657	2.026	10.780	0.000
Semantic	50000	1.002	0.995	1.009	0.709	0.502
	100000	0.999	0.988	1.010	-0.194	0.852
	150000	1.000	0.987	1.012	-0.089	0.932
	200000	0.994	0.987	1.002	-1.798	0.115
	250000	0.997	0.985	1.010	-0.512	0.625
	300000	0.996	0.980	1.011	-0.659	0.531
	350000	0.992	0.973	1.012	-0.936	0.380
	400000	1.006	0.976	1.037	0.478	0.648
	450000	1.078	0.994	1.163	2.195	0.064
	500000	1.160	1.042	1.279	3.207	0.015
	550000	1.231	1.113	1.349	4.623	0.002
	600000	1.298	1.167	1.429	5.372	0.001
	650000	1.345	1.172	1.518	4.721	0.002
	700000	1.395	1.191	1.600	4.569	0.003
	750000	1.501	1.269	1.734	5.101	0.001
	800000	1.521	1.300	1.741	5.587	0.001
	850000	1.602	1.321	1.882	5.071	0.001
	900000	1.525	1.243	1.807	4.400	0.003
	950000	1.562	1.313	1.812	5.325	0.001
	1000000	1.697	1.409	1.986	5.707	0.001

Table A4: Development of late competitor effects

Fine Grain

Competitor	Training Stage	Comp. Bias	Confidence Interval Lower	Upper	t-value	p-value
Phonological	50000	0.995	0.989	1.002	-1.787	0.117
	100000	0.994	0.987	1.001	-2.044	0.080
	150000	0.997	0.982	1.012	-0.431	0.680
	200000	1.008	0.996	1.020	1.510	0.175
	250000	1.000	0.982	1.018	0.004	0.997
	300000	1.005	0.979	1.032	0.484	0.644
	350000	1.013	0.986	1.039	1.128	0.297
	400000	1.049	1.016	1.083	3.477	0.010
	450000	1.140	1.059	1.220	4.097	0.005
	500000	1.122	1.026	1.218	3.001	0.020
	550000	1.173	1.034	1.312	2.944	0.022
	600000	1.223	1.049	1.397	3.032	0.019
	650000	1.139	0.971	1.306	1.959	0.091
	700000	1.156	0.992	1.319	2.248	0.059
	750000	1.139	0.961	1.317	1.849	0.107
	800000	1.178	0.952	1.404	1.865	0.105
	850000	1.166	0.936	1.395	1.703	0.132
	900000	1.152	0.962	1.342	1.887	0.101
	950000	1.245	0.985	1.505	2.231	0.061
	1000000	1.246	0.984	1.508	2.216	0.062
Visual	50000	1.001	0.992	1.010	0.318	0.760
	100000	1.000	0.990	1.009	-0.125	0.904
	150000	1.006	0.999	1.014	1.928	0.095
	200000	0.996	0.985	1.008	-0.784	0.459
	250000	1.020	1.002	1.037	2.691	0.031
	300000	1.005	0.981	1.029	0.447	0.668
	350000	1.023	0.998	1.048	2.141	0.070
	400000	1.064	1.017	1.111	3.200	0.015
	450000	1.188	1.083	1.294	4.217	0.004
	500000	1.455	1.269	1.642	5.763	0.001
	550000	1.730	1.496	1.964	7.381	0.000
	600000	1.827	1.587	2.067	8.152	0.000
	650000	1.809	1.599	2.018	9.118	0.000
	700000	1.898	1.597	2.200	7.035	0.000
	750000	2.006	1.792	2.220	11.109	0.000
	800000	2.078	1.852	2.305	11.269	0.000
	850000	2.039	1.761	2.317	8.837	0.000
	900000	2.049	1.782	2.315	9.302	0.000
	950000	2.087	1.767	2.407	8.035	0.000
	1000000	2.063	1.741	2.386	7.809	0.000
Semantic	50000	1.002	0.993	1.011	0.515	0.623
	100000	0.999	0.988	1.010	-0.217	0.834
	150000	1.001	0.987	1.014	0.098	0.925
	200000	0.994	0.985	1.003	-1.668	0.139
	250000	0.999	0.982	1.015	-0.221	0.832
	300000	0.992	0.970	1.014	-0.880	0.408
	350000	0.996	0.975	1.017	-0.503	0.631
	400000	1.032	1.010	1.054	3.487	0.010
	450000	1.174	1.055	1.294	3.462	0.011
	500000	1.401	1.187	1.615	4.431	0.003
	550000	1.667	1.363	1.971	5.188	0.001
	600000	1.742	1.503	1.981	7.334	0.000
	650000	1.770	1.425	2.116	5.276	0.001
	700000	1.864	1.529	2.199	6.095	0.001
	750000	2.028	1.671	2.385	6.810	0.000
	800000	1.981	1.656	2.306	7.136	0.000
	850000	2.084	1.644	2.523	5.832	0.001
	900000	1.900	1.519	2.281	5.592	0.001
	950000	1.986	1.623	2.348	6.429	0.000
	1000000	2.036	1.677	2.395	6.826	0.000

Table A5: ToW effects on eye gaze after 250k training trials

Fine Grain

Competitor	Time Step	Comp. Bias	Confidence Interval Lower	Confidence Interval Upper	t-value	p-value
Phonological	1	1.001	0.995	1.007	0.242	0.815
	2	0.998	0.986	1.011	-0.359	0.731
	3	0.997	0.975	1.018	-0.382	0.714
	4	0.995	0.966	1.023	-0.451	0.665
	5	0.992	0.957	1.028	-0.507	0.628
	6	0.990	0.951	1.030	-0.569	0.587
	7	0.989	0.945	1.034	-0.555	0.596
	8	0.987	0.937	1.037	-0.625	0.552
	9	0.985	0.929	1.041	-0.633	0.547
	10	0.983	0.922	1.045	-0.635	0.546
	11	0.981	0.914	1.048	-0.674	0.522
	12	0.979	0.907	1.051	-0.692	0.511
	13	0.976	0.899	1.052	-0.755	0.475
	14	0.972	0.893	1.052	-0.824	0.437
	15	0.970	0.887	1.053	-0.858	0.419
	16	0.967	0.882	1.051	-0.940	0.379
	17	0.965	0.878	1.051	-0.961	0.368
	18	0.963	0.874	1.052	-0.987	0.357
	19	0.961	0.871	1.050	-1.033	0.336
	20	0.960	0.868	1.052	-1.035	0.335
	21	0.959	0.865	1.053	-1.033	0.336
	22	0.960	0.865	1.054	-1.013	0.345
	23	0.958	0.864	1.052	-1.060	0.325
	24	0.958	0.863	1.054	-1.034	0.336
	25	0.958	0.861	1.055	-1.022	0.341
Visual	1	1.001	0.990	1.013	0.242	0.815
	2	0.996	0.972	1.021	-0.347	0.739
	3	0.994	0.952	1.035	-0.368	0.724
	4	0.991	0.935	1.046	-0.403	0.699
	5	0.987	0.918	1.056	-0.443	0.671
	6	0.984	0.907	1.060	-0.505	0.629
	7	0.982	0.895	1.068	-0.505	0.629
	8	0.975	0.878	1.073	-0.601	0.567
	9	0.969	0.860	1.078	-0.669	0.525
	10	0.965	0.845	1.085	-0.693	0.511
	11	0.960	0.829	1.091	-0.723	0.493
	12	0.954	0.815	1.094	-0.773	0.465
	13	0.948	0.800	1.095	-0.839	0.429
	14	0.941	0.788	1.094	-0.916	0.390
	15	0.935	0.776	1.094	-0.965	0.367
	16	0.930	0.767	1.093	-1.019	0.342
	17	0.925	0.759	1.091	-1.069	0.320
	18	0.921	0.750	1.091	-1.100	0.308
	19	0.918	0.746	1.089	-1.134	0.294
	20	0.915	0.739	1.091	-1.145	0.290
	21	0.913	0.734	1.091	-1.156	0.286
	22	0.912	0.731	1.092	-1.157	0.285
	23	0.910	0.729	1.091	-1.172	0.280
	24	0.909	0.727	1.092	-1.174	0.279
	25	0.910	0.724	1.095	-1.155	0.286
Semantic	1	1.001	0.990	1.013	0.242	0.815
	2	0.996	0.972	1.021	-0.347	0.739
	3	0.994	0.952	1.035	-0.368	0.724
	4	0.991	0.935	1.046	-0.403	0.699
	5	0.987	0.918	1.056	-0.443	0.671
	6	0.984	0.907	1.060	-0.505	0.629
	7	0.982	0.895	1.068	-0.505	0.629
	8	0.975	0.878	1.073	-0.601	0.567
	9	0.969	0.860	1.078	-0.669	0.525
	10	0.965	0.845	1.085	-0.693	0.511
	11	0.960	0.829	1.091	-0.723	0.493
	12	0.954	0.815	1.094	-0.773	0.465
	13	0.948	0.800	1.095	-0.839	0.429
	14	0.941	0.788	1.094	-0.916	0.390
	15	0.935	0.776	1.094	-0.965	0.367
	16	0.930	0.767	1.093	-1.019	0.342
	17	0.925	0.759	1.091	-1.069	0.320
	18	0.921	0.750	1.091	-1.100	0.308
	19	0.918	0.746	1.089	-1.134	0.294
	20	0.915	0.739	1.091	-1.145	0.290
	21	0.913	0.734	1.091	-1.156	0.286
	22	0.912	0.731	1.092	-1.157	0.285
	23	0.910	0.729	1.091	-1.172	0.280
	24	0.909	0.727	1.092	-1.174	0.279
	25	0.910	0.724	1.095	-1.155	0.286

Table A6: ToW effects on eye gaze after 500k training trials

Fine Grain

Competitor	Time Step	Comp. Bias	Confidence Interval Lower	Confidence Interval Upper	t-value	p-value
Phonological	1	0.998	0.994	1.001	-1.523	0.172
	2	1.000	0.979	1.021	-0.009	0.993
	3	1.002	0.953	1.051	0.082	0.937
	4	1.005	0.931	1.080	0.161	0.877
	5	1.008	0.913	1.104	0.203	0.845
	6	1.010	0.901	1.118	0.209	0.840
	7	1.018	0.899	1.138	0.362	0.728
	8	1.034	0.908	1.161	0.640	0.543
	9	1.070	0.932	1.207	1.196	0.271
	10	1.127	0.976	1.277	1.992	0.087
	11	1.206	1.038	1.373	2.901	0.023
	12	1.301	1.114	1.487	3.807	0.007
	13	1.391	1.187	1.595	4.536	0.003
	14	1.466	1.254	1.679	5.185	0.001
	15	1.518	1.302	1.733	5.685	0.001
	16	1.549	1.335	1.763	6.073	0.001
	17	1.556	1.349	1.763	6.349	0.000
	18	1.547	1.348	1.746	6.509	0.000
	19	1.526	1.335	1.717	6.519	0.000
	20	1.498	1.315	1.681	6.446	0.000
	21	1.466	1.289	1.643	6.218	0.000
	22	1.436	1.262	1.610	5.923	0.001
	23	1.410	1.237	1.582	5.617	0.001
	24	1.389	1.218	1.561	5.368	0.001
	25	1.375	1.204	1.546	5.195	0.001
Visual	1	0.995	0.988	1.003	-1.523	0.172
	2	0.991	0.968	1.014	-0.966	0.366
	3	0.987	0.940	1.033	-0.691	0.512
	4	0.986	0.920	1.051	-0.519	0.620
	5	0.986	0.909	1.062	-0.448	0.668
	6	0.986	0.906	1.065	-0.431	0.679
	7	0.987	0.908	1.066	-0.378	0.717
	8	0.990	0.918	1.063	-0.318	0.760
	9	1.006	0.931	1.081	0.198	0.849
	10	1.037	0.948	1.127	0.993	0.354
	11	1.078	0.962	1.194	1.590	0.156
	12	1.130	0.982	1.277	2.079	0.076
	13	1.193	1.017	1.369	2.586	0.036
	14	1.266	1.069	1.462	3.190	0.015
	15	1.345	1.135	1.555	3.879	0.006
	16	1.428	1.211	1.645	4.667	0.002
	17	1.507	1.289	1.725	5.503	0.001
	18	1.582	1.364	1.800	6.319	0.000
	19	1.648	1.429	1.867	6.990	0.000
	20	1.703	1.483	1.923	7.552	0.000
	21	1.743	1.521	1.964	7.934	0.000
	22	1.771	1.549	1.993	8.217	0.000
	23	1.790	1.568	2.011	8.413	0.000
	24	1.801	1.579	2.023	8.531	0.000
	25	1.804	1.582	2.026	8.570	0.000
Semantic	1	0.998	0.994	1.001	-1.523	0.172
	2	0.990	0.969	1.011	-1.097	0.309
	3	0.979	0.930	1.029	-0.997	0.352
	4	0.968	0.892	1.043	-1.020	0.342
	5	0.957	0.860	1.053	-1.063	0.323
	6	0.945	0.834	1.056	-1.165	0.282
	7	0.940	0.819	1.060	-1.188	0.274
	8	0.936	0.810	1.062	-1.204	0.268
	9	0.940	0.806	1.073	-1.065	0.322
	10	0.953	0.811	1.095	-0.779	0.462
	11	0.974	0.827	1.121	-0.421	0.686
	12	1.007	0.862	1.153	0.116	0.911
	13	1.054	0.915	1.194	0.917	0.390
	14	1.119	0.987	1.251	2.127	0.071
	15	1.194	1.066	1.322	3.582	0.009
	16	1.280	1.144	1.416	4.860	0.002
	17	1.365	1.212	1.518	5.647	0.001
	18	1.451	1.274	1.628	6.022	0.001
	19	1.529	1.330	1.727	6.304	0.000
	20	1.599	1.380	1.817	6.487	0.000
	21	1.655	1.421	1.889	6.617	0.000
	22	1.700	1.452	1.948	6.681	0.000
	23	1.734	1.477	1.992	6.736	0.000
	24	1.758	1.494	2.021	6.802	0.000
	25	1.771	1.505	2.037	6.850	0.000

Table A7: ToW effects on eye gaze after 750k training trials

Fine Grain

Competitor	Time Step	Comp. Bias	Confidence Interval Lower	Upper	t-value	p-value
Phonological	1	0.998	0.993	1.003	-0.899	0.399
	2	1.000	0.983	1.018	0.027	0.979
	3	1.008	0.973	1.043	0.535	0.609
	4	1.017	0.961	1.073	0.723	0.493
	5	1.027	0.950	1.105	0.825	0.436
	6	1.035	0.939	1.132	0.861	0.418
	7	1.041	0.931	1.152	0.885	0.406
	8	1.053	0.931	1.176	1.031	0.337
	9	1.080	0.956	1.203	1.528	0.171
	10	1.150	1.030	1.269	2.963	0.021
	11	1.303	1.180	1.427	5.807	0.001
	12	1.504	1.357	1.650	8.115	0.000
	13	1.686	1.507	1.865	9.060	0.000
	14	1.820	1.616	2.023	9.538	0.000
	15	1.900	1.683	2.117	9.805	0.000
	16	1.933	1.699	2.168	9.398	0.000
	17	1.920	1.666	2.175	8.550	0.000
	18	1.869	1.595	2.142	7.505	0.000
	19	1.789	1.512	2.066	6.745	0.000
	20	1.703	1.432	1.973	6.145	0.001
	21	1.619	1.366	1.872	5.783	0.001
	22	1.557	1.315	1.798	5.451	0.001
	23	1.525	1.289	1.761	5.264	0.001
	24	1.523	1.280	1.767	5.083	0.001
	25	1.553	1.299	1.808	5.148	0.001
Visual	1	0.996	0.985	1.007	-0.899	0.399
	2	0.981	0.959	1.003	-2.041	0.081
	3	0.971	0.927	1.016	-1.533	0.169
	4	0.966	0.898	1.035	-1.165	0.282
	5	0.965	0.876	1.054	-0.928	0.384
	6	0.965	0.863	1.067	-0.810	0.445
	7	0.963	0.851	1.074	-0.788	0.457
	8	0.961	0.847	1.075	-0.808	0.446
	9	0.963	0.856	1.070	-0.814	0.443
	10	0.978	0.881	1.075	-0.529	0.613
	11	1.022	0.925	1.118	0.532	0.611
	12	1.102	0.998	1.206	2.317	0.054
	13	1.221	1.109	1.334	4.652	0.002
	14	1.380	1.252	1.509	6.990	0.000
	15	1.579	1.414	1.743	8.314	0.000
	16	1.812	1.581	2.044	8.284	0.000
	17	2.055	1.736	2.375	7.810	0.000
	18	2.282	1.873	2.692	7.407	0.000
	19	2.466	1.992	2.939	7.325	0.000
	20	2.598	2.095	3.101	7.515	0.000
	21	2.663	2.163	3.163	7.870	0.000
	22	2.676	2.203	3.148	8.388	0.000
	23	2.646	2.212	3.080	8.975	0.000
	24	2.595	2.190	2.999	9.318	0.000
	25	2.544	2.163	2.926	9.576	0.000
Semantic	1	0.998	0.993	1.003	-0.899	0.399
	2	0.997	0.973	1.020	-0.330	0.751
	3	0.994	0.947	1.041	-0.301	0.772
	4	0.990	0.922	1.059	-0.337	0.746
	5	0.989	0.899	1.078	-0.303	0.771
	6	0.989	0.883	1.096	-0.238	0.819
	7	0.991	0.869	1.113	-0.175	0.866
	8	0.997	0.862	1.131	-0.060	0.954
	9	1.002	0.860	1.143	0.028	0.979
	10	1.008	0.863	1.154	0.137	0.895
	11	1.026	0.886	1.165	0.432	0.679
	12	1.061	0.934	1.188	1.134	0.294
	13	1.128	1.014	1.243	2.646	0.033
	14	1.236	1.118	1.354	4.746	0.002
	15	1.387	1.241	1.534	6.247	0.000
	16	1.578	1.378	1.778	6.840	0.000
	17	1.790	1.522	2.059	6.967	0.000
	18	2.002	1.661	2.342	6.959	0.000
	19	2.187	1.789	2.585	7.057	0.000
	20	2.337	1.893	2.780	7.128	0.000
	21	2.434	1.962	2.907	7.179	0.000
	22	2.483	1.995	2.970	7.193	0.000
	23	2.492	1.999	2.985	7.155	0.000
	24	2.471	1.975	2.968	7.006	0.000
	25	2.442	1.947	2.937	6.885	0.000

Table A8: ToW effects on eye gaze after 1m training trials

Fine Grain

Competitor	Time Step	Comp. Bias	Confidence Interval Lower	Upper	t-value	p-value
Phonological	1	0.998	0.991	1.004	-0.914	0.391
	2	1.004	0.981	1.028	0.434	0.678
	3	1.014	0.965	1.063	0.696	0.509
	4	1.029	0.954	1.103	0.902	0.397
	5	1.040	0.936	1.143	0.906	0.395
	6	1.050	0.913	1.186	0.860	0.419
	7	1.057	0.900	1.214	0.857	0.420
	8	1.066	0.902	1.230	0.953	0.372
	9	1.095	0.938	1.253	1.427	0.197
	10	1.198	1.051	1.344	3.187	0.015
	11	1.410	1.235	1.585	5.529	0.001
	12	1.664	1.446	1.881	7.209	0.000
	13	1.871	1.632	2.111	8.603	0.000
	14	2.015	1.763	2.268	9.511	0.000
	15	2.097	1.830	2.363	9.732	0.000
	16	2.121	1.835	2.407	9.266	0.000
	17	2.082	1.778	2.387	8.405	0.000
	18	1.999	1.688	2.309	7.610	0.000
	19	1.900	1.596	2.204	7.005	0.000
	20	1.804	1.511	2.096	6.486	0.000
	21	1.744	1.458	2.029	6.167	0.001
	22	1.723	1.437	2.008	5.989	0.001
	23	1.766	1.472	2.061	6.147	0.001
	24	1.837	1.541	2.132	6.696	0.000
	25	1.910	1.624	2.195	7.540	0.000
Visual	1	0.995	0.983	1.008	-0.914	0.391
	2	0.993	0.972	1.013	-0.860	0.418
	3	1.000	0.959	1.042	0.024	0.982
	4	1.014	0.947	1.081	0.480	0.646
	5	1.026	0.934	1.117	0.664	0.528
	6	1.035	0.924	1.145	0.745	0.481
	7	1.033	0.911	1.156	0.645	0.540
	8	1.033	0.909	1.156	0.625	0.552
	9	1.037	0.923	1.150	0.766	0.469
	10	1.061	0.958	1.163	1.399	0.205
	11	1.117	1.002	1.233	2.404	0.047
	12	1.225	1.082	1.368	3.726	0.007
	13	1.366	1.202	1.531	5.267	0.001
	14	1.547	1.341	1.752	6.296	0.000
	15	1.771	1.486	2.055	6.404	0.000
	16	2.035	1.633	2.438	6.086	0.001
	17	2.309	1.765	2.853	5.689	0.001
	18	2.551	1.893	3.210	5.572	0.001
	19	2.740	2.011	3.469	5.644	0.001
	20	2.833	2.088	3.577	5.821	0.001
	21	2.838	2.129	3.548	6.129	0.001
	22	2.763	2.116	3.410	6.444	0.000
	23	2.674	2.092	3.256	6.801	0.000
	24	2.598	2.065	3.131	7.087	0.000
	25	2.560	2.045	3.075	7.167	0.000
Semantic	1	0.998	0.991	1.004	-0.914	0.391
	2	1.004	0.986	1.023	0.533	0.611
	3	1.011	0.977	1.045	0.758	0.473
	4	1.017	0.962	1.071	0.728	0.490
	5	1.022	0.940	1.103	0.625	0.552
	6	1.029	0.919	1.140	0.626	0.551
	7	1.042	0.903	1.181	0.711	0.500
	8	1.053	0.893	1.214	0.787	0.457
	9	1.065	0.897	1.232	0.909	0.393
	10	1.073	0.908	1.238	1.048	0.329
	11	1.088	0.926	1.251	1.284	0.240
	12	1.121	0.976	1.265	1.974	0.089
	13	1.182	1.065	1.299	3.685	0.008
	14	1.289	1.182	1.396	6.383	0.000
	15	1.447	1.313	1.582	7.862	0.000
	16	1.658	1.460	1.856	7.854	0.000
	17	1.906	1.620	2.192	7.490	0.000
	18	2.157	1.788	2.526	7.419	0.000
	19	2.383	1.952	2.813	7.599	0.000
	20	2.539	2.075	3.003	7.843	0.000
	21	2.618	2.147	3.089	8.121	0.000
	22	2.603	2.148	3.057	8.332	0.000
	23	2.554	2.124	2.985	8.532	0.000
	24	2.492	2.086	2.897	8.699	0.000
	25	2.437	2.049	2.825	8.752	0.000

Table A9: ToW effects with reduced visual preview

Full Visual Acquity = word onset							Full Visual Acquity = word onset + 10						
Competitor	Time Step*	Comp. Bias	Confidence Interval Lower	Upper	t-value	p-value	Competitor	Time Step*	Comp. Bias	Confidence Interval Lower	Upper	t-value	p-value
Phonological	1	1.004	0.950	1.057	0.188	0.860	Phonological	1	1.000	1.000	1.000	NA	NA
	2	1.004	0.945	1.064	0.203	0.849		2	1.004	0.970	1.038	0.317	0.767
	3	1.118	1.025	1.210	3.536	0.024		3	1.009	0.950	1.068	0.419	0.697
	4	1.265	1.115	1.414	4.919	0.008		4	1.014	0.941	1.086	0.525	0.627
	5	1.407	1.191	1.622	5.241	0.006		5	1.019	0.945	1.093	0.700	0.522
	6	1.521	1.241	1.801	5.159	0.007		6	1.026	0.957	1.095	1.037	0.358
	7	1.593	1.250	1.936	4.803	0.009		7	1.033	0.971	1.094	1.480	0.213
	8	1.618	1.224	2.013	4.348	0.012		8	1.037	0.986	1.088	2.031	0.112
	9	1.607	1.180	2.035	3.942	0.017		9	1.040	0.995	1.085	2.449	0.071
	10	1.574	1.130	2.017	3.592	0.023		10	1.041	0.996	1.085	2.533	0.065
	11	1.524	1.071	1.978	3.213	0.033		11	1.040	0.996	1.083	2.541	0.064
	12	1.476	1.016	1.937	2.871	0.045		12	1.065	1.014	1.115	3.544	0.024
	13	1.440	0.978	1.902	2.644	0.057		13	1.115	1.033	1.197	3.877	0.018
	14	1.416	0.963	1.868	2.552	0.063		14	1.188	1.048	1.327	3.722	0.020
	15	1.411	0.969	1.853	2.581	0.061		15	1.266	1.050	1.481	3.424	0.027
	16	1.412	0.976	1.847	2.625	0.059		16	1.334	1.037	1.632	3.122	0.036
	17	1.418	0.983	1.853	2.666	0.056		17	1.388	1.018	1.758	2.913	0.044
	18	1.424	0.988	1.859	2.701	0.054		18	1.425	1.003	1.847	2.794	0.049
	19	1.426	0.993	1.860	2.730	0.053		19	1.446	1.002	1.889	2.788	0.049
	20	1.419	1.002	1.835	2.791	0.049		20	1.447	1.010	1.884	2.838	0.047
Visual	1	0.995	0.936	1.054	-0.231	0.829	Visual	1	1.000	1.000	1.000	NA	NA
	2	0.996	0.929	1.063	-0.170	0.873		2	1.017	0.993	1.042	1.934	0.125
	3	1.074	1.002	1.147	2.846	0.047		3	1.029	0.992	1.066	2.180	0.095
	4	1.210	1.110	1.310	5.807	0.004		4	1.038	0.989	1.088	2.171	0.096
	5	1.390	1.232	1.548	6.867	0.002		5	1.042	0.989	1.094	2.205	0.092
	6	1.600	1.369	1.831	7.203	0.002		6	1.043	0.992	1.095	2.333	0.080
	7	1.823	1.507	2.138	7.244	0.002		7	1.042	0.995	1.089	2.498	0.067
	8	2.043	1.641	2.445	7.202	0.002		8	1.040	0.996	1.084	2.500	0.067
	9	2.247	1.761	2.733	7.129	0.002		9	1.035	0.990	1.080	2.144	0.099
	10	2.428	1.880	2.976	7.233	0.002		10	1.028	0.982	1.075	1.682	0.168
	11	2.571	1.965	3.176	7.202	0.002		11	1.024	0.975	1.072	1.358	0.246
	12	2.674	2.015	3.333	7.051	0.002		12	1.108	1.088	1.128	15.204	0.000
	13	2.736	2.042	3.429	6.951	0.002		13	1.289	1.231	1.348	13.738	0.000
	14	2.751	2.053	3.449	6.962	0.002		14	1.563	1.392	1.735	9.137	0.001
	15	2.727	2.043	3.411	7.009	0.002		15	1.872	1.564	2.179	7.874	0.001
	16	2.668	1.996	3.339	6.892	0.002		16	2.161	1.711	2.610	7.173	0.002
	17	2.599	1.927	3.271	6.611	0.003		17	2.398	1.821	2.975	6.725	0.003
	18	2.542	1.875	3.208	6.424	0.003		18	2.577	1.899	3.255	6.459	0.003
	19	2.508	1.845	3.171	6.314	0.003		19	2.691	1.965	3.416	6.472	0.003
	20	2.480	1.836	3.124	6.384	0.003		20	2.738	2.007	3.470	6.599	0.003
Semantic	1	1.003	0.973	1.033	0.268	0.802	Semantic	1	1.000	1.000	1.000	NA	NA
	2	1.004	0.972	1.036	0.349	0.745		2	1.008	0.982	1.034	0.876	0.431
	3	1.055	1.014	1.096	3.742	0.020		3	1.011	0.970	1.052	0.735	0.503
	4	1.143	1.089	1.198	7.300	0.002		4	1.015	0.968	1.062	0.892	0.423
	5	1.266	1.193	1.338	10.225	0.001		5	1.020	0.976	1.063	1.255	0.278
	6	1.417	1.311	1.523	10.952	0.000		6	1.025	0.988	1.063	1.871	0.135
	7	1.585	1.417	1.753	9.669	0.001		7	1.026	0.994	1.058	2.271	0.086
	8	1.752	1.504	2.000	8.421	0.001		8	1.026	0.996	1.057	2.383	0.076
	9	1.913	1.575	2.251	7.492	0.002		9	1.026	0.989	1.063	1.959	0.122
	10	2.064	1.641	2.487	6.980	0.002		10	1.028	0.982	1.073	1.674	0.170
	11	2.194	1.697	2.691	6.669	0.003		11	1.031	0.980	1.082	1.685	0.167
	12	2.295	1.744	2.847	6.516	0.003		12	1.101	1.056	1.147	6.154	0.004
	13	2.365	1.782	2.948	6.499	0.003		13	1.233	1.171	1.295	10.376	0.001
	14	2.399	1.809	2.989	6.582	0.003		14	1.430	1.311	1.548	10.055	0.001
	15	2.405	1.811	2.999	6.566	0.003		15	1.655	1.447	1.864	8.716	0.001
	16	2.379	1.776	2.982	6.346	0.003		16	1.870	1.551	2.190	7.563	0.002
	17	2.337	1.726	2.948	6.072	0.004		17	2.051	1.618	2.484	6.739	0.003
	18	2.291	1.674	2.908	5.811	0.004		18	2.195	1.657	2.734	6.165	0.004
	19	2.253	1.627	2.879	5.558	0.005		19	2.298	1.698	2.898	6.006	0.004
	20	2.215	1.591	2.838	5.410	0.006		20	2.350	1.724	2.975	5.993	0.004

* Time steps record number of time steps post word onset (e.g. 1 = word onset +1)

Table A10: ToW effects on semantic activation after 1m training trials

Fine Grain

Competitor	Time Step	Dist.	Confidence Interval		t-value	p-value
			Lower	Upper		
Phonological	1	1.005	1.000	1.011	2.384	0.049
	2	1.023	1.011	1.035	4.387	0.003
	3	1.032	1.013	1.052	3.933	0.006
	4	1.035	1.012	1.058	3.528	0.010
	5	1.035	1.007	1.063	2.995	0.020
	6	1.033	1.001	1.066	2.435	0.045
	7	0.996	0.958	1.034	-0.251	0.809
	8	0.840	0.785	0.896	-6.770	0.000
	9	0.666	0.601	0.730	-12.219	0.000
	10	0.706	0.656	0.755	-14.097	0.000
	11	0.810	0.774	0.845	-12.763	0.000
	12	0.885	0.857	0.914	-9.631	0.000
	13	0.928	0.903	0.953	-6.768	0.000
	14	0.953	0.929	0.977	-4.661	0.002
	15	0.967	0.944	0.991	-3.239	0.014
	16	0.977	0.953	1.001	-2.281	0.057
	17	0.983	0.959	1.008	-1.638	0.146
	18	0.987	0.962	1.012	-1.211	0.265
	19	0.990	0.965	1.016	-0.902	0.397
	20	0.993	0.968	1.018	-0.688	0.514
	21	0.994	0.969	1.019	-0.534	0.610
	22	0.996	0.971	1.020	-0.415	0.691
	23	0.997	0.973	1.021	-0.318	0.760
	24	0.998	0.974	1.021	-0.241	0.816
	25	0.998	0.975	1.022	-0.174	0.867
Visual	1	1.002	0.995	1.009	0.741	0.483
	2	1.019	1.011	1.027	5.646	0.001
	3	1.026	1.014	1.038	5.036	0.002
	4	1.028	1.014	1.043	4.737	0.002
	5	1.030	1.014	1.045	4.439	0.003
	6	1.031	1.014	1.048	4.221	0.004
	7	1.036	1.016	1.057	4.199	0.004
	8	1.037	1.014	1.060	3.825	0.007
	9	1.030	1.005	1.055	2.865	0.024
	10	1.018	0.995	1.042	1.820	0.112
	11	1.009	0.988	1.029	0.980	0.360
	12	1.004	0.985	1.022	0.497	0.634
	13	1.002	0.984	1.019	0.216	0.835
	14	1.001	0.984	1.017	0.077	0.941
	15	1.000	0.984	1.016	-0.028	0.978
	16	0.999	0.984	1.015	-0.105	0.919
	17	0.999	0.984	1.014	-0.163	0.875
	18	0.999	0.984	1.014	-0.197	0.849
	19	0.999	0.984	1.013	-0.221	0.831
	20	0.999	0.985	1.012	-0.251	0.809
	21	0.998	0.985	1.012	-0.285	0.784
	22	0.998	0.986	1.011	-0.309	0.766
	23	0.998	0.986	1.011	-0.317	0.761
	24	0.999	0.987	1.010	-0.308	0.767
	25	0.999	0.988	1.010	-0.274	0.792
Semantic	1	1.001	0.992	1.010	0.268	0.797
	2	1.011	0.994	1.029	1.586	0.157
	3	1.017	0.991	1.044	1.545	0.166
	4	1.016	0.985	1.047	1.233	0.258
	5	1.012	0.978	1.045	0.827	0.436
	6	1.006	0.970	1.041	0.366	0.725
	7	0.981	0.941	1.022	-1.111	0.303
	8	0.895	0.849	0.942	-5.329	0.001
	9	0.809	0.769	0.849	-11.350	0.000
	10	0.680	0.654	0.706	-28.854	0.000
	11	0.589	0.574	0.604	-63.778	0.000
	12	0.554	0.541	0.568	-76.621	0.000
	13	0.545	0.530	0.559	-73.808	0.000
	14	0.542	0.527	0.557	-71.600	0.000
	15	0.542	0.526	0.557	-70.203	0.000
	16	0.542	0.526	0.558	-68.407	0.000
	17	0.543	0.526	0.559	-66.730	0.000
	18	0.543	0.527	0.560	-65.032	0.000
	19	0.544	0.527	0.561	-63.105	0.000
	20	0.545	0.527	0.563	-61.257	0.000
	21	0.546	0.528	0.564	-59.736	0.000
	22	0.546	0.528	0.564	-58.496	0.000
	23	0.547	0.528	0.565	-57.463	0.000
	24	0.547	0.528	0.566	-56.438	0.000
	25	0.549	0.530	0.568	-56.266	0.000
Target	1	0.997	0.992	1.003	-1.138	0.293
	2	1.022	0.998	1.046	2.149	0.069
	3	1.033	1.000	1.066	2.356	0.051
	4	1.037	1.000	1.073	2.351	0.051
	5	1.038	0.998	1.077	2.255	0.059
	6	1.037	0.995	1.080	2.066	0.078
	7	0.996	0.945	1.048	-0.171	0.869
	8	0.838	0.778	0.899	-6.284	0.000
	9	0.666	0.614	0.718	-15.238	0.000
	10	0.404	0.366	0.442	-37.134	0.000
	11	0.215	0.188	0.243	-68.091	0.000
	12	0.132	0.114	0.151	-110.020	0.000
	13	0.101	0.086	0.117	-134.838	0.000
	14	0.089	0.074	0.104	-143.901	0.000
	15	0.084	0.070	0.099	-146.259	0.000
	16	0.083	0.068	0.098	-145.390	0.000
	17	0.083	0.067	0.098	-142.507	0.000
	18	0.083	0.068	0.099	-138.916	0.000
	19	0.085	0.069	0.101	-136.025	0.000
	20	0.087	0.071	0.103	-134.199	0.000
	21	0.090	0.074	0.106	-133.240	0.000
	22	0.093	0.077	0.109	-132.319	0.000
	23	0.097	0.080	0.113	-130.570	0.000
	24	0.101	0.084	0.117	-127.399	0.000
	25	0.106	0.088	0.123	-122.658	0.000

Chapter 4

The multimodal nature of spoken word processing in the visual world: Testing the predictions of a multimodal integration model (MIM)[1]

Abstract

Ambiguity in natural language is ubiquitous, yet spoken communication is effective due to integration of information carried in the speech signal with information available in the surrounding multimodal landscape. Language mediated visual attention requires visual and linguistic information integration and has thus been used to examine properties of the architecture supporting multimodal processing during spoken language comprehension. In this paper we implemented parallel, multimodal processing explicitly in a computational model that combines phonological, semantic and visual processing information streams. The model generated novel predictions about stronger and earlier influences of visual and semantic similarity compared to phonological similarity around the rhyme of words, which were confirmed in two visual world studies. During spoken word comprehension, multimodal information can be recruited rapidly to constrain processing to the extent that phonological rhyme information may often exert little influence on this process.

[1] *Smith, A. C., Monaghan, P., & Huettig, F. (submitted): The multimodal nature of spoken word processing in the visual world: Testing the predictions of a multimodal integration model (MIM)*

1. Introduction

One of the defining features of language is displacement, i.e., the fact that concepts need not refer to objects or events that are currently present (Hockett & Altmann, 1968). In line with this observation is a long tradition of research in the language sciences which has largely ignored potential influences of 'non-linguistic' information sources (e.g., Fodor, 1983). However, although language does not need to refer to objects which are physically present it is often used in such a way. Moreover, psycholinguistic research over recent years suggests that language processing (including spoken word processing) is highly interactive in terms of combining multiple information sources to form an interpretation of the signal. It is therefore likely to be a profound misrepresentation to restrict models of spoken word recognition exclusively to auditory information, overlooking multimodal aspects of the speech processing system (e.g. McClelland & Elman, 1986; Luce et al., 2000 Norris & McQueen, 2008; Scharenborg & Boves, 2010).

Indeed, the prevalence of ambiguity in natural language (Piantadosi, Tily & Gibson, 2012) is evidence for the efficiency with which the human speech processing system integrates linguistic and extra-linguistic information. If we accept that language usage takes place in context (i.e., embedded within extra-linguistic factors, world context, world knowledge, etc.) then the amount of information an efficient language should convey must be less than the amount of information required out of context (Monaghan, Christiansen, & Fitneva, 2011; Piantadosi et al., 2012). To take the extreme case, if the language system operated independently of context then an efficient language cannot contain ambiguity. However, we know ambiguity in natural language is ubiquitous yet such ambiguity is rarely harmful to effective communication (Piantadosi et al., 2012; Wasow & Arnold, 2003; Wasow et al., 2005; Jaeger, 2006; Roland, Elman & Ferreira, 2006; Ferreira, 2008; Jaeger, 2010). This implies that the speech processing system is able to efficiently integrate extra-linguistic contextual information with the ambiguous speech stream it receives. The lack of explicit awareness we have of the level of ambiguity within the raw speech signal when processing speech in natural settings illustrates the speed and ease with which linguistic and non-linguistic information is integrated by the human speech processing system.

Within this paper we frame spoken word recognition and spoken word comprehension in terms of multimodal constraint satisfaction (cf. MacDonald et al., 1994; McClelland, Rumelhart, & Hinton, 1986; McClelland et al., 2014). Words can be conceived as entities that

connect representations across multiple modalities (e.g. phonological, orthographic, semantic, visual, etc). Thus, human speech processing occurs in a multimodal context, with activation of information across modalities in constant flux reflecting real time sensory input computed through past experience and current cognitive constraints. An efficient spoken word recognition system should therefore rapidly incorporate such multimodal cues. This then allows the system to adapt its responses in accordance to the current multimodal evidential landscape. It may of course be possible for a message to be independent of the immediate non-auditory sensory environment, however an efficient system must also be sensitive to cues in the combined multimodal signal that would allow it to make this distinction.

Models of speech recognition and speech comprehension have frequently overlooked this multimodal aspect of the speech processing system (e.g., Luce et al., 2000; McClelland & Elman, 1986; Norris & McQueen, 2008; Scharenborg & Boves, 2010), with comparatively little known about the architecture that supports integration and the temporal structure of this process. Here we provide an explicit description of an architecture able to support multimodal language processing that considers spoken word recognition in terms of multimodal constraint satisfaction. We provide simulations of how visual, phonological and semantic information may be integrated when processing spoken words in a visual world. In two eye-tracking experiments we probe participants' abilities to integrate visual and linguistic information in order to test whether such an architecture accurately predicts participants' gaze when presented with objects that share either visual, semantic, or phonological properties with a spoken target word.

Visual world eye-tracking as a method to study spoken word processing

Visual world experiments, in which participants' gaze is recorded when mapping between visual and auditory stimuli, have been used extensively to examine the interface between visual and linguistic processing streams (see Huettig, Rommers & Meyer, 2011, for a recent review). These studies detail a number of word-level language-mediated eye-gaze phenomena that have been used to infer properties of the speech processing system. Specifically, they provide insight into the type of information activated as a spoken word unfolds, the relative influence of specific sources of information during speech comprehension, and the temporal structure of this process. Such inferences are based on the assumption that gaze towards an item reflects the level to which properties of the item

(*relative to all other items within the display*) are activated at a given point in time by the speech signal.

We know from such studies that items in the visual environment whose names share their phonological onset with a spoken target word (e.g., beaver and beaker) can attract visual attention from shortly after word onset (Allopenna et al. 1998). We also know from the same study that visually displayed items whose names share their phonological rhyme with the spoken target word (e.g., speaker and beaker) are also fixated more than unrelated objects shortly after target word onset, yet slightly later than objects that share their phonological onsets. But it is not only the activation of phonological information that has been indexed by such studies of language mediated visual attention. They have also demonstrated that items that share visual properties (e.g., shape: beaker and bobbin) with a spoken target word (but no phonological relationship) attract attention early in post word onset (Dahan & Tanenhaus, 2005; Huettig & Altmann, 2007; Huettig & McQueen, 2007). Items that share semantic (but not phonological or visual) relationships with spoken target words (e.g., cent and purse) also have been demonstrated to attract attention rapidly post word onset (Dunabeitia, el al., 2009; Huettig & Altmann, 2005; Huettig et al. 2006; Yee & Sedivy, 2006). Together, these data demonstrate that as a spoken word unfolds, its phonological, visual and semantic properties are activated rapidly and thus can be recruited to map onto information extracted from the immediate visual environment.

To examine the relative timing of activation of phonological, semantic and visual information by the unfolding speech signal, Huettig and McQueen (2007) presented participants with scenes containing items that shared properties of the target word in one of each of these three dimensions. Scenes contained an item which shared its phonological onset with the spoken target word (phonological onset competitor); an item that shared visual properties with the spoken target word (visual competitor); an item that shared semantic properties with the spoken target word (semantic competitor); and an item that was unrelated to the spoken word in all three dimensions (unrelated distractor). They observed that participants first looked towards phonological competitors while later looking towards visual and semantic competitors once later phonemes had provided disambiguating information to discount the phonological competitor. This pattern of gaze was interpreted by Huettig & McQueen as evidence for the cascaded activation of information through the speech processing system, with the speech signal first activating the target word's phonological properties, then later visual and semantic properties. The approach of pairing multiple items within the visual

display that contrast in the properties they share with the spoken target word, in order to expose differences in the temporal structure of activation of such properties during spoken word processing, was also used in Huettig and Altmann (2011), to demonstrate earlier activation of semantic properties compared to an item's surface colour. Using the same visual world paradigm they observed that participants looked first to items that shared semantic properties of the spoken target word (e.g. spinach – mushroom) then later to items that shared the (prototypical) surface colour of the spoken target word (e.g. spinach – green dress).

Similarly, pairing items within the visual display that contrast in the properties they share with the spoken target word has also been used to examine the relative influence of a given property on language mediated eye gaze, and, by extension, motivate statements regarding relative activation during spoken word processing. Allopenna et al. (1998) presented scenes containing items that either shared their phonological onset or rhyme with the spoken target word. They observed that participants' gaze towards phonological rhyme competitors occurred later and was weaker than onset effects. Studies of rhyme competitor effects have since shown that they result in typically only small, marginally significant effects (see also Allopenna et al., 1998; McQueen & Huettig, 2012; McQueen & Viebahn, 2007). This indicates that phonological information in the onset is more influential in spoken word recognition than information carried in the rhyme.

The use of language mediated eye gaze to make statements about spoken word recognition has gained influence due to a coupling of visual world data and computational models of spoken word recognition. This approach requires the explicit description of the mechanisms driving eye gaze that can be tested against behavioural findings. Allopenna et al.'s observation of an influence of rhyme competitors on fixation behaviour proved notable as this was initially believed to be a point of distinction between alternative models of spoken word recognition: such as continuous mapping models (e.g. TRACE: McClelland & Elman, 1986) and alignment models (e.g. Marslen-Wilson, 1987; Norris, 1990). In early descriptions of alignment models, initial phonemes constrain the candidate set of words such that words that mismatched at onset, such as rhyme competitors, are no longer under consideration. Hence, should such an alignment model be driving fixation behaviour, then fixation of rhyme competitors should not exceed levels displayed towards unrelated items. In contrast, within continuous mapping models, mapping occurs across the entire word with overall similarity driving words' level of activation. Thus, words that share their rhyme, yet not their onset, will still be activated. TRACE, the continuous mapping model used in Allopenna et al. (1998),

predicts both a rhyme effect, and also a distinction in the level of activation of onset and rhyme competitors. As onset phonemes are encountered earlier, their activation will, before the overlapping phonemes in the rhyme unfold, inhibit rhyme competitors. Hence, TRACE predicts that rhyme competitors will be activated at levels lower than those of onset competitors, which was the pattern observed in Allopenna et al. (1998). Although continuous mapping models had predicted the influence of phonological rhyme overlap during spoken word recognition, evidence for such an influence had been difficult to isolate using standard priming paradigms (Andruski, Blumstein, & Burton, 1994; Connine et al., 1993). Eye gaze in the visual world paradigm, however, offers a temporally rich measure that provided the necessary sensitivity to potentially capture these subtle effects (Allopenna et al., 1998).

It has since been demonstrated that alignment models are also capable of generating rhyme competitor effects if they are exposed to noise in the learning environment, such that onset information is not always a perfect predictor of the target word (Magnuson, Tanenhaus & Aslin, 2000; Magnuson et al., 2003; Smith, Monaghan & Huettig, 2013). Evidence to support such predictions is provided by recent visual world data that demonstrates that onset and rhyme effects on language mediated eye gaze can be modulated by the level of noise participants are exposed to in the speech signal (McQueen & Huettig, 2012).

In sum, studies of language mediated visual attention have demonstrated that visually displayed items that share their phonological rhyme with the spoken target word attract attention more than unrelated items. However, such effects have been small and generally only marginally significant. Further, such effects have only been observed under heavily controlled laboratory conditions, in which phonology is the only property connecting items in the display to the spoken target word. Therefore, it remains an open question whether phonological rhyme information exerts an influence on language mediated eye gaze when other sources of information are available to map between visual and auditory streams. Such data also offers an indirect measure of the relative influence of phonological rhyme information during day-to-day spoken word processing, in situations when information from semantic or visual modalities may also be available to constrain spoken word recognition and comprehension.

The Multimodal Integration Model (MIM) of language -mediated visual attention

A distinct division in perspectives continues to exist within both cognitive psychology and cognitive neuroscience regarding the characterisation of how and when non-linguistic and

linguistic information interact during speech processing (see Pulvermuller et al., 2009, for a review). Late interaction models (Fodor, 1983; Friederici, 2002; Marslen-Wilson, 1987; Morton, 1969; Shallice, 1988) characterise processing in terms of a modular, serial or slowly cascading system. Within such models, stored knowledge relating to a word is accessed via a fixed sequence of discrete stages of processing, often with hundreds of milliseconds required to initiate each subsequent stage and in which the activation of stored knowledge is initially driven purely by the phonetic properties of the spoken word. In contrast, within early interaction models (Gaskell & Marslen-Wilson, 1997, 2002; MacDonald et al., 1994; Pulvermuller et al., 2009; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) multiple forms of information (e.g., phonological, syntactic, semantic, visual, etc.) contribute early and at roughly the same time, approximating parallel processing. Within such models representations emerge from a continuous process in which multiple forms of information interact in parallel, rather than discrete stages of processing leading to a point of selection of an item from a distinct candidate set.

Table 1: Table presents data recorded in the Visual World Paradigm that the MIM model has previously been demonstrated to capture (Smith, Monaghan & Huettig, 2013, 2014)

Study	**Scene**			
Authors (year)	**Item 1**	**Item 2**	**Item 3**	**Item 4**
Allopenna et al. (1998)	Target	**Onset**	**Rhyme**	Distractor
Dahan & Tanenhaus (2005)	Target	**Visual**	Distractor	Distractor
Huettig & Altmann (2007)	**Visual**	Distractor	Distractor	Distractor
Yee & Sedivy (2006)	Target	**Semantic**	Distractor	Distractor
Huettig & Altmann (2005)	**Semantic**	Distractor	Distractor	Distractor
Mirman & Magnuson (2009)[a]	Target	**Near Semantic**	**Far Semantic**	Distractor
Huettig & McQueen (2007)[b]	**Onset**	**Semantic**	**Visual**	Distractor

Notes: Item 1-4 indicate the relationship of each of the four objects presented in the visual display of each study to the spoken target word. Observed competitor effects are indicated in bold type. Onset = phonological onset competitor; Rhyme = phonological rhyme competitor; Visual = visual competitor; Semantic = semantic competitor.

[a] *Near and Far semantic competitors presented on separate trials*

[b] *Experiment 1*

As previously described, language mediated eye gaze offers a temporally rich measure that can expose subtle properties of processing which, through extracting predictions from explicit implementation in computational models, can be used to test theoretical models of the architecture of the speech processing system. Most studies exploring word level effects with this approach have focused on how language-mediated visual attention can be explained purely in terms of phonological processing (Allopenna et al., 1998; Magnuson et al., 2003; McMurray et al., 2010; Mirman et al., 2008). The literature outlined above demonstrates that language mediated eye gaze is dependent on the interaction of phonological, visual and semantic information, it therefore offers a novel means of examining how such sources of information may interact when mapping between visual and linguistic streams which can also be used to make inferences regarding the architecture supporting multimodal integration during spoken word processing. In order to answer questions relating to, for example, the relative influence of phonological rhyme, semantic and visual information on processing, it is necessary to provide a model in which all modalities are able to interact. Multimodal models have been used in conjunction with visual world data (Kukona & Tabor, 2011; Mayberry et al., 2009; Mirman & Magnuson, 2009; Smith, Monaghan & Huettig, 2013, 2014; Spivey, 2007), however only the recently developed multimodal integration model (MIM, Smith et al., 2013) combines the following properties of language mediated eye gaze that together have allowed it to capture a comprehensive range of word level effects reported in the visual world literature (see Table 1): competition at multiple levels of representation in semantic, visual and phonological dimensions; parallel activation of representations; and integration of visual, semantic and phonological processing streams. We describe the model in detail below.

The Multimodal Integration Model (MIM) frames spoken word recognition and comprehension in terms of multimodal constraint satisfaction, offering a parsimonious solution to how visual, semantic and phonological information are integrated during spoken word processing. The model proposes that concurrent phonological, semantic and visual information are integrated in parallel during spoken word processing and as a computational model it provides an explicit description of how this may be implemented in a cognitively plausible architecture (McClelland, Mirman, Bolger & Khaitan, 2014).

Aims of the current study

Our aim is to examine the interaction of phonological rhyme, semantic and visual information within language mediated visual attention. These data will then be used to motivate

inferences regarding the architecture supporting such multimodal interaction during spoken word processing. This is achieved by first using the MIM model to generate predictions for how gaze should be distributed towards visual, semantic and phonological rhyme competitors given that concurrent visual, semantic and phonological information is processed in parallel during spoken word recognition. Two visual world experiments will then measure behaviour of participants when exposed to the conditions simulated in the model, in order to evaluate model predictions.

The first visual world experiment presented scenes that contained a single phonological rhyme competitor and three unrelated distractors. This will establish whether relationships within the materials are sufficient to generate the rhyme effect reported in previous visual world studies. A second visual world experiment presented the same scenes as used in the first experiment yet with two of the unrelated distractors replaced with a visual and a semantic competitor. The second experiment thus examined how the phonological rhyme effect is affected by competition from semantic and visual competitors.

A comparison between experiment 1 (rhyme competitor only) and experiment 2 (rhyme, semantic and visual competitors) offers four possible outcomes: 1) there is no difference in the rhyme effect observed in experiment 1 and 2, therefore rhyme effects are not altered by the presence of visual and semantic competitors; 2) the rhyme effect is present in both experiments yet the presence of visual and semantic competitors weakens the rhyme effect; 3) the rhyme effect is present in both experiments yet the presence of visual and semantic competitors increases the rhyme effect; or 4) the rhyme effect is only present in experiment 1, therefore the presence of semantic and visual competitors eliminates the rhyme effect.

These four outcomes vary in their compatibility with late versus early integration models. Serial processing models would predict no influence of increased competition on phonological rhyme effects. Should processing proceed serially with the word's full phonological form processed prior to visual and semantic activation, then rhyme effects should be unaffected by the additional competition provided by visual and semantic competitors as such properties will not become active until after rhyme information has had an opportunity to influence gaze.

Cascade models vary in their predictions depending on the speed with which information transfers between levels of representation. If the system slowly cascades information from the early activation of a word's phonological form to later activation of its visual and semantic

properties then rhyme effects should either show little influence of the increased competition in experiment 2 (outcome 1) or display an increasing influence of competition at later stages of processing (outcome 2).

Rapid cascade or parallel processing models are potentially compatible with all outcomes. Should information cascade rapidly or in parallel to activate information in all three dimensions then it is possible that rhyme effects will be observed at a similar level in experiments 1 and 2. However, in such circumstances visual and semantic effects should also be observed within the same or earlier time windows as such information will be available to map between visual and auditory processing streams and hence be able to influence fixations in these time windows. With this additional mapping a possibility within the system, the second or fourth outcome is more likely in a parallel processor, because eye gaze can only be attributed to a single object at any point in time. Gaze should then be drawn towards visual and semantic competitors within the same time window in which phonological rhyme overlap exerts an effect, meaning that the proportion of fixations towards rhyme competitors will be lower and hence a reduced rhyme effect would be observed in experiment 2 compared to experiment 1.

The following section provides a brief overview of the multimodal parallel integration model and the simulations of experiment 1 and 2 conducted using the model. This is then followed by a description of two experimental visual world studies. Results of the simulations are then evaluated in light of experimental findings and their consequences for language mediated eye gaze research and, more broadly, spoken word processing.

2. Simulating the effects of multimodal competition on phonological rhyme overlap in language mediated visual attention

Model

The Multimodal Integration Model (Smith, Monaghan, & Huettig, 2013) of language mediated visual attention was used for simulations within this study. The architecture, representations and training procedure replicated that described in Smith, Monaghan & Huettig, (2013)[1]. An overview of the implementation is provided below, for a full description

[1] Model used within this study replicates the 'noisy learning environment' implementation described within Smith, Monaghan & Huettig, (2013).

of the motivation for and structure of the model, refer to Smith et al. (2013). The model aims to provide a parsimonious and explicit description of the information and mechanisms driving language mediated visual attention and how behaviour can emerge from the statistical constraints of the learning environment. Previous studies have demonstrated the model's ability to capture a broad range of word level properties of language mediated visual attention (see Table 1).

Architecture

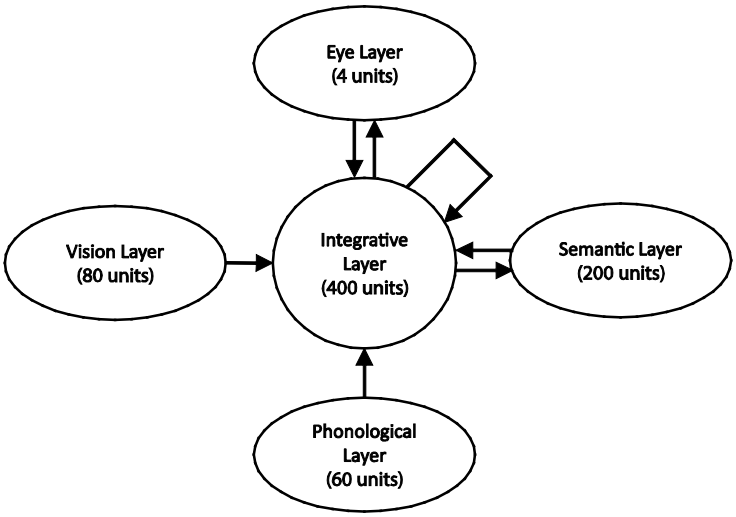

Figure 1: Architecture of the multimodal parallel integration *model of language mediated visual attention.*

The MIM model utilises the parallel distributed processing framework (see Rumelhart, McClelland & the PDP Research Group, 1986; Rogers & McClelland, 2014). The network consists of layers of non-linear processing units connected via weighted connections (see appendix). The architecture of the model is displayed in Figure 1. A layer of 80 units defines the visual layer. This layer provides input of visual information to the network from four locations (each represented by 20 units) in the visual field. A layer of 60 units provides input of phonological information to the network. This layer is divided into six phoneme slots, with each slot consisting of 10 units each sensitive to a specific phonological property of an utterance at a specific temporal location. Units in both phonological and visual layers are fully connected in a forward direction to a central integrative layer. The integrative layer consisted of 400 units and is fully self-connected. The integrative layer is also fully connected to both a semantic layer and an eye layer, in both forward and backward

directions. The semantic layer consists of 200 units each of which are sensitive to a specific semantic property. The eye layer consists of four units with each unit encoding the probability that the model directs gaze to one of the four locations in the visual field.

Representations

Table 2: Details of relationships between targets, competitors and unrelated distractors embedded within artificial corpora.

Representation	**Item Type**	**Constraint (Features shared with target)**	**Cosine Distance**
Phonological	Competitor	Final 3 of 6 phonemes	0.259
	Unrelated	Max. 2 consecutive phonemes	0.496
Semantic	Competitor	4 of 8 semantic features	0.500
	Unrelated	Max. 1 semantic feature	0.959
Visual	Competitor	Min. 5 of 10 visual features	0.264
	Unrelated	Features shared with p = (0.5)	0.506

Eight artificial corpora were constructed, with each used to train and test a single simulation run of the model, this ensured that relationships within and between modalities were controlled. Each corpus consisted of 200 words, with each word assigned a unique phonological, semantic and visual representation. All words within the corpus were six phonemes in length. A phoneme inventory consisting of 20 phonemes was constructed, with each phoneme represented by a unique 10 unit binary phonological feature vector. Each phonological feature was assigned with p(active) = 0.5. Phonological representations were constructed by pseudo randomly sampling a unique sequence of six phonemes from the phoneme inventory to create each word. Controls ensured no more than 2 consecutive phonemes were shared between words (other than in the case of phonological onset or rhyme competitors, see Table 2). Visual representations were unique 20 unit binary feature vectors, with each unit representing the presence or absence of a specific visual feature. Visual features were assigned with p(active) = 0.5. Semantic representations by contrast were sparsely distributed, with each word pseudo-randomly assigned a unique set of eight semantic features from a possible 200. A maximum of 1 semantic property was shared between items (other than in the case of semantic competitors where 4 properties were shared, see Table 2).

Embedded within the corpus were 20 sets of items that shared increased overlap in either semantic, visual or phonological dimensions. Each of the 20 sets contained a target, a phonological competitor, a semantic competitor and a visual competitor. Each 'target' word shared the final three of its six phonemes with the phonological rhyme competitor, four of its eight semantic features with the semantic competitor and a minimum of 5 of its 10 visual features with the visual competitor. This ensured that in all dimensions the distance between competitor and target was half that between competitor and an unrelated item (see Table 2).

Training

In training the model we assume that individuals learn associations between representations of an item across modalities through repeated, simultaneous exposure to multiple representational forms of an item. Networks were trained on four cross modal mapping tasks: object recognition; spoken word comprehension; speech motivated orientation; and semantically motivated orientation. Time in the model was represented by the flow of information across weighted connections between units in the network (see appendix). Each training task ran for 14 time steps (ts).

For object recognition tasks, four items were randomly selected from the training corpus and their visual representations presented to the four visual input slots within the visual layer (ts =0). One of the four items was then randomly selected as a target and the eye gaze layer unit corresponding to the location of the target's visual representation in the visual layer was fully activated (ts =0). Visual input and eye gaze layer activation remained fixed across the training trial while random time invariant noise was provided as an input to the phonological layer. At ts 3 until the end of the trial the semantic representation of the target was presented to the semantic layer and error back propagated.

Spoken word comprehension tasks involved randomly selecting an item from the corpus as a target. The phonological representation of the target item was then over time presented to the phonological layer of the network, with an additional phoneme presented at each subsequent time step. To simulate exposure to noise in the auditory input within the learning environment the binary value of each unit within the phonological representation of the target was switched (i.e. 0 -> 1 or 1 -> 0) with p = 0.2 (see Smith, Monaghan & Huettig, 2013). Random time invariant noise was presented as input to the visual layer during such trials, while no constraints were placed on eye layer activity. At ts 5 the semantic representation of

the target was presented to the semantic layer and error backpropagated until the end of the training trial.

For phonological orientation tasks, four items were randomly selected and their visual representations presented as input to the visual layer (ts 0 – 14). One of the four items was randomly selected as a target. The target's phonological representation was then presented over time as input to the phonological layer, with an additional phoneme presented at each subsequent time step. As in word comprehension tasks, to simulate exposure to noisy auditory signals in the learning environment the value of each unit in the target's phonological representation was switched with p = 0.2. No constraints were placed on activity in the semantic layer. At ts 5 (point of phonological disambiguation) the eye layer unit corresponding to the location of the target's visual representation was required to be fully activated and error backpropagated.

Finally, semantic orientation trials followed a similar procedure. Again four items were randomly selected from the corpus and their visual representations presented as input to the visual layer (ts 0-14). One of these four items was randomly selected as a target and its semantic representation presented to the semantic layer (ts 0-14). Random time invariant noise was presented to the phonological layer throughout this trial. At ts 2 the eye layer unit that corresponded to the location of the visual representation of the target was required to be fully activated and error backpropagated.

We assume that speech motivated orientation is less frequent in the learning environment than object recognition, spoken word comprehension and semantically motivated orientation and therefore this task was four times less likely to occur during training. Given this constraint training tasks were randomly interleaved. Connection weights within the model were initialised with random weights from the uniform distribution [-0.1, 0.1]. Recurrent backpropagation (learning rate = 0.05) was used during training to adjust weights within the network (see appendix). A total of 125000 training trials were performed before the model was exposed to test conditions. Once trained all networks performed spoken word comprehension and object recognition tasks accurately (i.e. semantic layer activity was closest in terms of cosine distance to that of the target) for all items in the training corpus. On orientation tasks the model looked to the location of the target on at least 3 of 4 test trials for 99.75% (speech motivated orientation) and 100% (semantically motivated orientation) of items. Eight simulation runs of the model were run each initiated with a different initial

random seed. Results report mean behaviour calculated across all eight simulation runs of the model.

Simulation 1: Simulating effects of phonological rhyme overlap

Previous visual world studies demonstrate that phonological rhyme overlap exerts an influence on language mediated visual attention under conditions in which phonology provides the only dimension in which auditory and visually presented stimuli are related. We first examine the model's sensitivity to phonological rhyme overlap when presented with scenes containing a single rhyme competitor and three unrelated items.

Procedure

The visual representations of four objects were presented to the visual layer at trial onset and remained so until the end of the trial (ts 0-30). Three of the items were unrelated to the upcoming target word, i.e., controlled low level of overlap with the target in visual, semantic or phonological dimensions (see Table 2). The fourth item was a phonological rhyme competitor in that it shared the final three phonemes of its phonological representation with the upcoming target word. The network was then free to cycle for five time steps to allow pre-processing of the visual information. At ts 5 the phonological representation of the target word began to unfold with an additional phoneme presented at each subsequent time step. Unlike in training, no noise was applied to the phonological input of the target representation. Activation in the eye layer was recorded throughout the trial. The location in the visual field fixated by the model at a given time point was recorded as the location associated with the most activated unit in the eye layer at the given point in time. This procedure was followed for all rhyme competitor and target pairs within the corpus (n = 20) with rhyme competitors and distractors tested in all possible combinations of location (n = 24) resulting in a total of 480 test trials per simulation run of the model.

Results

Figure 2 presents the change in the probability of fixating rhyme competitors and unrelated distractors from word onset, with the probability of fixating unrelated distractors divided by three as three were presented in the display alongside a single rhyme competitor. To examine whether looks to phonological rhyme competitors exceeded looks to unrelated distractors for analysis we divided the 30 time step (ts) test trial into six equal time windows (ts 1 – 5; ts 6 – 10; ts 11 – 15; ts 16 – 20; ts 21 – 25; ts 26 - 30). We then compared fixation behaviour displayed by the model in the time window prior to word onset (ts 1 – 5) to fixation

behaviour displayed by the model in each of the five time windows post word onset. For each window we calculated the empirical log odds of fixating each category of object within the display (i.e., rhyme competitor, unrelated distractor). This measure avoids issues arising from calculating estimates based on proportional data (see Jaeger, 2008). Our dependent measure was the difference between the log-odds of fixating the phonological rhyme competitor and the log-odds of fixating the unrelated distractor. This reflects the difference in fixation of competitor objects as a consequence representational overlap. We used linear mixed effect models to examine whether gaze differed as a consequence of phonological rhyme overlap in the time windows post word onset relative to levels of fixation prior to word onset. The model constructed applied the maximal random effect structure (Barr, Levy, Scheepers, & Tily, 2013), the fixed effect time window and random effects of model simulation run (n = 8) and item (n = 20), including random intercepts and slopes for time window both by subject and item. To derive p-values we assumed t-values were drawn from a normal distribution (Barr, 2008).

Examining parameter estimates within the model revealed that in the first time block that followed word onset (ts 6 - 10) phonological rhyme competitors were fixated slightly less than unrelated distractors relative to the period prior to word onset (β = -0.190, t = -2.086, p = 0.037). In the second time block (ts 11 – 15), there was no difference in fixation of rhyme competitors compared to unrelated distractors (β = 0.075, t = 1.031, p = 0.302). While in time windows ts 16 – 20 (β = 0.367, t = 2.73, p = 0.006), ts 21 – 25 (β = 0.300, t = 2.479, p = 0.013) and ts 26 – 30 (β = 0.304, t = 2.324, p = 0.02), phonological rhyme competitors were fixated more that unrelated distractors.

$$Competitor\ Bias = \frac{probability\ of\ fixating\ competitor}{\sum(probability\ of\ fixating\ unrelated\ distractor\ /\ number\ of\ unrelated\ distractors)} \quad (1)$$

To examine the relative magnitude of the effect and the onset of phonological rhyme effects we used t-tests to examine whether fixation of phonological rhyme competitors differed from distractors at each time step within the test trial relative to levels at word onset. We calculated the change in competitor bias (see equation 1) from word onset and compared at each time step whether this value differed from zero (see Appendix table 1). An increase from word onset (i.e. ΔCompetitor Bias > 0) would indicate increased fixation of the competitor compared to unrelated items post word onset, while a decrease from word onset (ΔCompetitor Bias < 0) would indicate increased fixation of unrelated distractors compared

to competitors relative to word onset levels. This analysis shows that the model first fixates phonological rhyme competitors above unrelated distractor levels from time step 8 - 9 (by simulation run, ts 8: M = 0.159, CI [0.005 0.312], t(7) = 2.439, p = 0.045; by item, ts = 9: M = 0.323, CI [0.073 0.574], t(19) = 2.705, p = 0.014). Prior to time step 8 – 9 there was no difference between fixation of rhyme competitors and unrelated distractors ($t < 2$, $p > 0.05$). Once the rhyme effect emerges phonological rhyme competitors remain fixated above unrelated distractor levels for all remaining time steps ($t > 3$, $p < 0.01$). The phonological rhyme effect was greatest at time step 12 (by item: M = 0.594, CI [0.212 0.976], t(19) = 3.256, p = 0.004; by simulation run: M = 0.546, CI [0.302 0.790], t(7) = 5.289, p = 0.001).

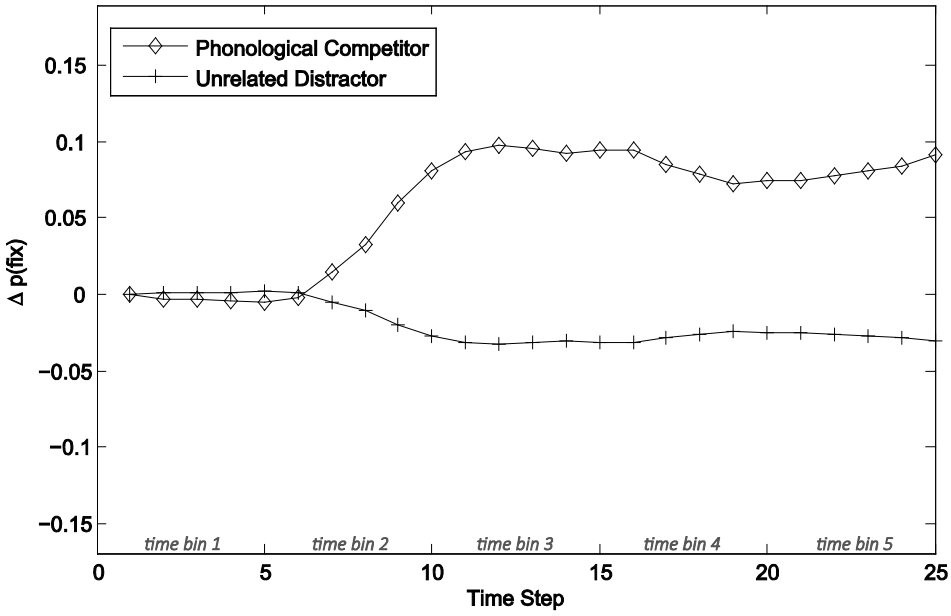

Figure 2: Change in proportion of fixations from word onset displayed by the multimodal parallel integration model to items in visual displays containing a rhyme competitor and three unrelated distractors (mean fixation of unrelated distractors plotted).

Discussion

Results of simulation 1 demonstrate that the MIM model displays sensitivity to phonological rhyme overlap when presented with scenes containing a single rhyme competitor amongst unrelated items. This replicates previous behavioural findings that language mediated eye gaze is sensitive to phonological rhyme overlap between spoken target words and visually displayed objects (Allopenna et al., 1998; Huettig & McQueen, 2012; McQueen & Viebahn, 2007). Further, the model demonstrates that an alignment model of spoken word processing

is able to generate phonological rhyme effects (Magnuson, Tanenhaus & Aslin, 2000; Magnuson et al., 2003; Smith, Monaghan & Huettig, 2013).

Simulation 2: Simulating effects of multimodal competition

A second set of simulations examined the relative influence and timing of effects of phonological rhyme, semantic and visual overlap on eye gaze within the MIM model and the effect of additional competition from visual and semantic competitors on phonological rhyme effects that were exhibited in Simulation 1.

Procedure

Simulation 2 followed the same training and testing procedure as outlined for Simulation 1, however test scenes now contained a rhyme competitor, a semantic competitor, a visual competitor and an unrelated distractor. Again simulations were run for all target and competitor sets embedded within the corpus (n = 20) with sets tested in all possible combinations of location (n=24) resulting in a total of 480 test trials per simulation run. Results report the probability of fixating an item at any given time point, this is taken as the proportion of trials on which at that given point in the trial the eye layer unit associated with location of the given object is the most activated unit in the eye layer.

Results

The change in the probability of fixating each category of item (i.e., rhyme competitor, semantic competitor, visual competitor and unrelated distractor) from word onset is presented in Figure 3. Visual inspection suggests a rapid increase in fixation of visual competitors shortly after word onset, with increased fixation of semantic competitors emerging later and at lower levels. Fixation of phonological rhyme competitors also appears to depart from unrelated distractor levels however this occurs later than semantic and visual competitors and is a weaker effect.

We used the same procedure for analysis of Simulation 2 as used in Simulation 1. However, scenes in Simulation 2 contained three competitors rather than a single competitor in simulation 1. We therefore compared separately for each category of competitor (visual, semantic, rhyme) the difference in empirical log odds of fixating the given competitor and the unrelated distractor in each time window post word onset (ts 6 – 10, ts 11 – 15, ts 16 – 20, ts 21 – 25, ts 26 - 30) to the difference observed in the period prior to word onset (ts 1 – 5). For analysis we used linear mixed effect models with a fixed effect of time window and random

effects of model simulation run (n = 8) and item (n = 20), including random intercepts and slopes for time window both by subject and item.

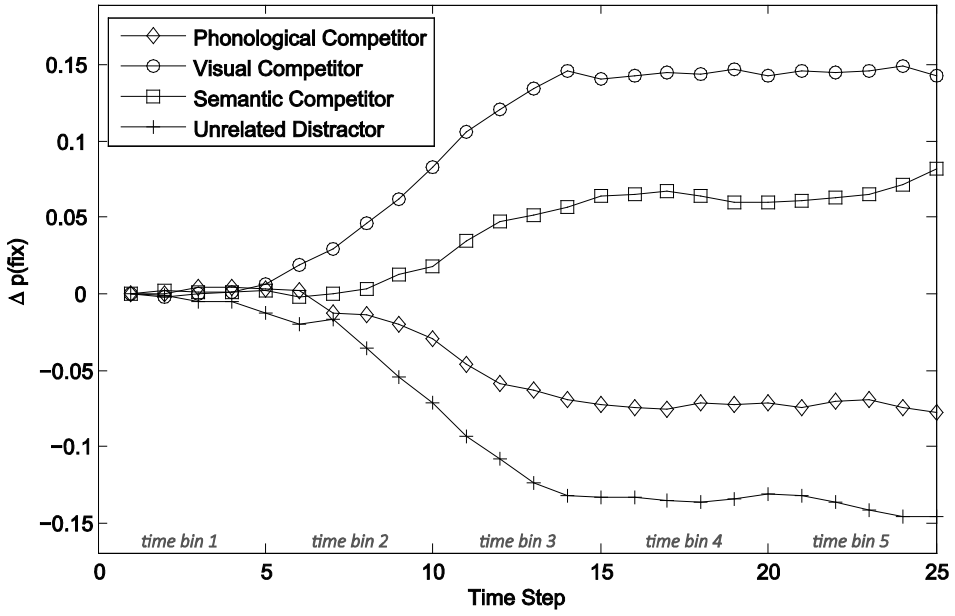

Figure 3: Change in the proportion of fixations from word onset displayed by the multimodal parallel integration *model to items in visual displays containing a rhyme competitor, a visual competitor, a semantic competitor and an unrelated distractor.*

This analysis revealed that visual and semantic competitors were fixated above unrelated distractor levels in windows ts 11 – 15 (visual: β = 0.262, t = 2.567, p = 0.01; semantic: β = 0.232, t = 2.164, p = 0.03), ts 16 – 20 (visual: β = 0.950, t = 7.074, p = < 0.001; semantic: β = 0.781, t = 5.124, p = < 0.001), ts 21 – 25 (visual: β = 1.118, t = 7.797, p = < 0.001; semantic: β = 0.938, t = 6.379, p = < 0.001) and ts 26 – 30 (visual: β = 1.157, t = 8.731, p = < 0.001; semantic: β = 0.989, t = 5.646, p = < 0.001). Although, in the first time block (ts 6 – 10) there was no difference between semantic competitors and unrelated distractors (β = 0.038, t = 0.287, p = 0.774), nor visual competitors and unrelated distractors (β = -0.109, t = -0.919, p = 0.358). Effects of phonological rhyme overlap were later to emerge, with no effect observed in the first (β = 0.018, t = 0.137, p = 0.891), or the second time block (β = 0.110, t = 0.861, p = 0.389). There was a marginal effect of phonological rhyme overlap in the third time block (β = 0.258, t = 1.896, p = 0.058). However, phonological rhyme competitors were fixated more than unrelated distractors in the time blocks 4 (β = 0.300, t = 2.531, p = 0.011) and 5 (β = 0.330, t = 3.008, p = 0.003).

Using the same analysis technique we examined whether the log odds of fixating the visual competitor differed from the log odds of fixating the semantic competitor in each of the post word onset time windows relative to differences present prior to word onset. This analysis did not reveal a difference in fixation of semantic competitor compared to fixation of visual competitors in any of these time blocks (1: $\beta = -0.148$, $t = -1.113$, $p = 0.266$; 2: $\beta = 0.0302$, $t = 0.278$, $p = 0.781$; 3: $\beta = 0.168$, $t = 1.034$, $p = 0.301$; 4: $\beta = 0.180$, $t = 1.105$, $p = 0.269$; 5: $\beta = 0.169$, $t = 0.974$, $p = 0.33$).

As with results of Simulation 1, in order to identify the relative onsets of competitor effects and their relative magnitudes at each time step we examined whether the competitor bias (see equation 1) differed from word onset levels for each category of competitor (visual, semantic, rhyme) at each individual time step post word onset. Using t-tests competitor bias relative to word onset was calculated at each time step post word onset and compared to zero both by item ($n = 20$) and by simulation run ($n = 8$) (see appendix table 2). This revealed that visual effects were first to emerge between time steps 6 – 8 (by item, ts 6: $M = 0.166$, CI [0.007 0.324], $t(19) = 2.183$, $p = 0.042$; by simulation run, ts 8: $M = 0.388$, CI [0.040 0.736], $t(7) = 2.632$, $p = 0.034$), remaining above unrelated distractor levels for all remaining time steps ($t > 2.10$, $p < 0.05$) and peaking in magnitude between time step 21 – 24 (by item, ts 21: $M = 3.350$, CI [1.578 5.122], $t(19) = 3.956$, $p < 0.001$; by simulation run, ts 24: $M = 2.553$, CI [1.902 3.204], $t(7) = 9.273$, $p < 0.001$). Semantic effects emerged between time steps 8 – 10 (by item, ts = 8: $M = 0.200$, CI [0.015 0.385], $t(19) = 2.271$, $p = 0.035$; by simulation run, ts = 10: $M = 0.587$, CI [0.123 1.051], $t(7) = 2.990$, $p = 0.020$), also remaining above distractor levels for all remaining time steps ($t > 2.10$, $p < 0.05$) other than time step 10 at which point the effect was marginal when comparing by item ($M = 0.583$, CI [-0.007 1.172], $t(19) = 2.068$, $p = 0.053$). Semantic effects peaked between time steps 21 – 25 (by item, ts = 21: $M = 3.558$, CI [0.074 7.042], $t(19) = 2.138$, $p = 0.046$; by simulation run, ts = 25: $M = 2.258$, CI [1.418 3.098], $t(7) = 6.355$, $p < 0.001$). Stable phonological rhyme effects were later to emerge between time steps 13 – 17 (by item, ts = 17: $M = 0.483$, CI [0.100 0.866], $t(19) = 2.640$, $p = 0.016$; by simulation run, ts 13: $M = 0.471$, CI [0.043 0.900], $t(7) = 2.600$, $p = 0.035$), these effects remained present for the majority of remaining time steps ($t > 2.1$, $p < 0.05$) with the exception being time step 21 (by item: $M = 1.072$, CI [-0.302 2.446], $t(19) = 1.633$, $p = 0.120$; by simulation run: $M = 0.623$, CI [-0.095 1.341], $t(7) = 2.051$, $p = 0.080$). However, a small effect of phonological rhyme overlap was also present at time step 6 when analysing by item ($M = 0.163$, CI [0.004 0.321], $t(19) = 2.150$, $p = 0.045$). Marginal effects

of phonological rhyme overlap were also present in the by item analysis at time steps 3, 5 and 13 – 16 ($2.15 > t > 1.90$, $0.10 > p > 0.05$).

Finally, using mixed effects models we analysed whether the difference in empirical log odds of fixating the phonological rhyme competitor and empirical log odds of fixating the unrelated distractor differed between simulation 1 and simulation 2. Did the presence of visual or semantic competitors influence the magnitude of the phonological rhyme effect? This was performed using a model with fixed effects of time window (pre word onset, post word onset: with the pre word onset time window mapped onto the intercept forming the baseline condition) and simulation (Simulation 1, Simulation 2: with Simulation 1 mapped onto the intercept forming the baseline condition) and random effects of model simulation run (n = 8) and item (n = 20), including random intercepts and slopes for time window and simulation both by subject and item. This analysis revealed no significant interaction between time window and simulation for any time window post word onset (ts 6 – 10: $\beta = 0.208$, $t = 1.477$, $p = 0.140$; ts 11 – 15: $\beta = 0.035$, $t = 0.301$, $p = 0.764$; ts 16 – 20: $\beta = -0.108$, $t = -0.828$, $p = 0.408$; ts 21 – 25: $\beta = 0.00001$, $t = 0.000$, $p = 1.000$; ts 26 – 30: $\beta = 0.026$, $t = 0.195$, $p = 0.845$), therefore there was no difference in magnitude of phonological rhyme effects displayed by the model in Simulation 1 compared to Simulation 2.

Discussion

Simulations with the multimodal integration model (MIM) predict that participants' gaze should be drawn towards items in the immediate visual environment that share visual properties, semantic properties, or the phonological rhyme of the spoken target word more than items that are unrelated in these three dimensions. The model also predicts that the onset and the magnitude of visual and semantic competitor effects should precede and exceed the onset and the magnitude of phonological rhyme competitor effects.

Visual effects within the model emerged earlier than phonological rhyme effects because early (in addition to late) phonemes belonging to the spoken target word are associated with the visual properties of the visual competitor provided as a direct input throughout the test trial. In contrast, it is only later phonemes in the rhyme of the spoken target word that are associated with the visual properties of the phonological rhyme competitor, therefore these phonemes only influence fixations at a later point in time. The semantic effect emerges at a similar time to that of the visual competitor effect yet earlier than the phonological rhyme effect. Within the model, early phonemes activate the associated semantic properties of the

target word, and because some of these semantic properties are shared with the semantic competitor these properties are also associated to the visual properties of the semantic competitor, thus fixations towards the semantic competitor increase. Again, rhyme effects can only emerge once later overlapping phonemes in the rhyme are presented and hence rhyme effects also emerge later than semantic effects. The short delay in semantic effects compared to visual effects is likely due to the time needed for the target word's phonological properties to activate corresponding semantic properties that drive fixation of semantic competitors. In contrast visual properties are provided as a direct input and therefore they can immediately influence fixation behaviour once associated phonological features of the target word are presented. Any potential differences in the magnitude of semantic and visual effects are also likely due to this contrast in the factors driving these two distinct effects. The magnitude of semantic effects is dependent on the level to which training results in the activation of semantic features which are shared between semantic competitor and target, whereas visual properties shared between visual competitor and target are always fully activated due to being provided as a direct input.

Visual and semantic effects increased over the course of the trial as more of the phonology of the target word unfolds, because this increases the number of active phonological properties associated with the visual properties of the visual competitor and the number of semantic properties associated with the visual properties of the semantic competitor. The continued unfolding of phonological properties of the spoken target word also increases the phonological rhyme effect, however unlike the visual and semantic competitors only phonological features in the rhyme have associations with properties of the phonological rhyme competitor. Due to the temporal structure of phonological representations these associations are likely weaker than those activated by features in the onset and therefore phonological rhyme effects are weaker than visual or semantic effects. This is the result of later phonemes contributing less to the identification of the target during training due to the system often having sufficient information to identify the target before they become available.

A comparison between Simulations 1 and 2 generates the prediction that the presence of visual and semantic competitors should delay the onset of phonological rhyme effects but should not affect the overall magnitude of the phonological rhyme effect observed.

3. Testing the effects of multimodal competition on phonological rhyme overlap in the visual world paradigm

The predictions of the model were next tested in two visual world experiments that exposed participants to the same experimental conditions as were examined in Simulations 1 and 2.

Experiment 1: Effects of phonological rhyme overlap in target absent scenes

In Experiment 1 participants were presented with scenes containing four items while hearing a spoken target word. On experimental trials a single item within the display shares its phonological rhyme with the spoken target word and was the only relationship to exist between these two stimuli, with the remaining three items unrelated in visual, semantic and phonological dimensions.

Participants

40 participants (mean age = 21.6 years, range 18 – 30 years) recruited from the MPI for Psycholinguistics subject database were paid for participation in this study. All were native speakers of Dutch, had no known hearing problems and had corrected or normal vision.

Materials

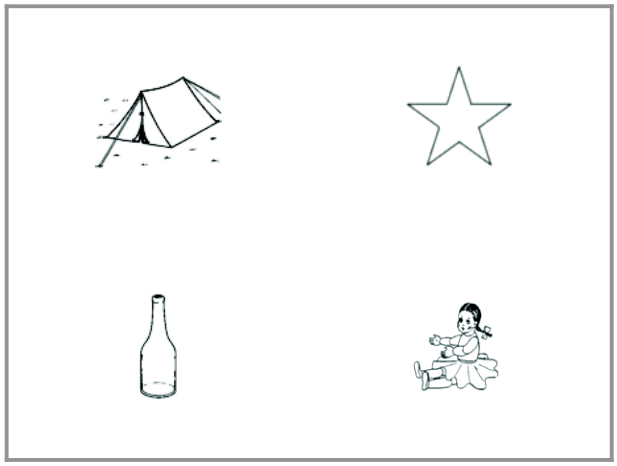

Figure 4: Example of experimental display from experiment 1. Within this trial the spoken target word was "cent", the rhyme competitor was 'tent' this is accompanied by three unrelated distractors 'pop' (doll), 'ster' (star) and 'fles' (bottle).

15 experimental trials and 34 filler trails were constructed, each consisting of a visual display and spoken Dutch sentence. Each sentence consisted of a target word embedded in a neutral carrier sentence in which the target word was not predictable (e.g. Dutch: "Zij begrepen niet

waarom de roos verwelkt was", English Translation: "They could not understand why the rose was withered"). Approximately six words (Experimental trials: mean = 6.33, SD = 1.53; Filler Trials: 6.90, SD = 1.58) preceded the target word in the carrier sentences. Spoken sentences were recorded in a sound dampened room by a female native Dutch speaker who was not aware of the purpose of the study. Instruction was provided for sentences to be read in a neutral tone and to avoid highlighting individual words within the sentence.

Table 3: Properties [μ (σ)] of words within competitor sets. Frequency = word frequency; Letters = number of letters; Syllables = number of syllables; Phonemes = number of phonemes shared with target spoken word; Semantic = semantic similarity rating to spoken target word; Visual = visual similarity rating to spoken target word.

Exp.	Item	Frequency	Letters	Syllables	Phonemes	Semantic	Visual
1	Rhyme	17.6	4.20	1.13	2.60	1.37	1.31
		(23.24)	(0.41)	(0.35)	(0.63)	(0.46)	(0.84)
	Dist.1	27.8	4.73	1.27	0.33	1.36	1.33
		(32.16)	(1.16)	(0.46)	(0.49)	(0.53)	(0.88)
	Dist.2	23.9	4.13	1.07	0.47	1.65	1.65
		(33.88)	(0.35)	(0.26)	(0.83)	(0.78)	(0.97)
	Dist.3	35.0	4.27	1.27	0.33	1.41	1.42
		(45.50)	(0.46)	(0.46)	(0.49)	(0.69)	(1.26)
2	Rhyme	17.6	4.20	1.13	2.60	0.95	1.23
		(23.24)	(0.41)	(0.35)	(0.63)	(0.46)	(0.74)
	Sem.	21.5	4.67	1.33	0.47	5.90	2.30
		(26.37)	(1.23)	(0.62)	(0.64)	(1.42)	(1.03)
	Visual	31.13	4.40	1.20	0.33	1.78	6.51
		(81.82)	(0.91)	(0.41)	(0.62)	(0.89)	(1.13)
	Dist.	30.5	4.67	1.27	0.27	1.36	1.53
		(32.06)	(1.23)	(0.46)	(0.46)	(0.67)	(0.90)

Visual displays contained black and white line drawings of four objects. Each object was resized to fit an area 96 x 96 pixels. The four images were presented in the four corners of the 1024 x 768 pixel display (locations: 256 x 192; 256 x 576; 768 x 192; 768 x 578). Seventeen (target present) filler trials, two of which were used as practice trials, contained an image of

the target word accompanied by three unrelated distractors. Seventeen (target absent) filler trials, two of which were used as practice trials, contained four unrelated distractors. Fifteen experimental trials contained a phonological rhyme competitor accompanied by three unrelated distractors (see Figure 4).

Word frequency, number of letters and number of syllables were controlled between competitor and unrelated distractor sets (see Table 3). All target, phonological rhyme competitor and unrelated distractor words were monosyllabic (see Appendix for a full list of experimental items). Phonological rhyme competitors were defined by the fact that they only differed from the target word in their initial phoneme (mean shared phonemes = 2.6, SD = 0.63). Controls ensured no sequence of phonemes was shared between target words and unrelated distractors. Separate semantic and visual similarity rating studies were conducted to ensure visual and semantic similarity was controlled across competitor and distractor sets.

Thirteen native Dutch speaking participants provided visual similarity ratings and a different group of 11 native Dutch speaking participants provided semantic similarity ratings, none of these participants later completed either of the eye tracking studies. Ratings were acquired using an online experiment in which participants were presented with the written form of the target word and the images corresponding to the rhyme competitor and distractors. In the case of visual similarity ratings participants were required to provide for each image a value between 0 and 10 indicating how similar the physical shape of the item in the image was to objects they associate with the target word (0 indicating no similarity in physical shape and 10 indicating both items have an identical physical shape), while ignoring any other relationships between the items for example semantic relationships. Similarly for semantic similarity ratings, for each image participants provided a value between 0 and 10 indicating how much of the target word's meaning is shared with the item depicted (0 indicating no similarity in meaning and 10 indicating complete overlap in meaning), while ignoring any other relationships between the items for example similarities in their physical shape. Results of these norming studies show that rhyme competitor and distractor sets did not differ in their semantic or visual similarity ratings to the spoken target words.

To ensure that the names attributed to displayed images were well motivated a picture name correspondence pre-test was conducted. 13 native Dutch speakers participated in this norming study and did not participate in either eye tracking experiment. Each image from experimental displays was presented to participants accompanied by either its intended name

or a randomly selected name. Participants were required to indicate whether the name corresponded to the image or not. Of the 60 words tested 52 were rated as corresponding by 100% of participants, 6 words by 92%, 1 word (Dutch: *vest*; English: *waistcoat*) by 85% and 1 word (Dutch: *kennel;* English: *kennel*) by 75%.

Procedure

An Eyelink 1000 tower mounted eye tracker (sampling rate 1kHz) was used to record participants' eye movements as they viewed displays on a computer monitor and listened to sentences through headphones while in a sound dampened room. Stimuli were presented and data recorded using the SR-Research program Experiment builder.

Participants performed a 'look-and-listen' task (see Huettig et al., 2011, for further discussion), they were instructed to look at the screen while listening carefully to sentences they would hear through the headphones. Trials followed the same procedure as reported in Huettig and McQueen (2007). The experimenter initiated the start of each trial when the participant fixated a fixation cross in the centre of the screen, this allowed for drift correction in the calibration if required between trials. Once the trial was initiated the fixation cross remained in the centre of the screen for 500ms, this was followed by a blank screen for 600ms. Then a scene containing four images was presented with display onset coinciding with the onset of the spoken sentence. The scene remained displayed for 4300ms (length of longest spoken sentence), following which a blank screen was presented for 500ms after which the trial ended. Eye gaze was recorded at all stages of the trial. The location of targets and competitors was randomised across trials, while the location of items and order of trials was randomised across participants. Before the experiment began each participant first completed four practice trials (2 x target present filler trials, 2 x target absent filler trials). In total the experiment lasted approximately 15 minutes.

Results

Four interest areas were defined for each experimental display that covered the area 270 x 235 pixels that surrounded each image within the scene. A fixation was recorded as directed towards an item if it fell within the interest area within which the given item was situated. Blinks and saccades were not included in the analysis. Figure 5 displays a time-course graph on which the difference in the proportion of fixations from target word onset directed towards rhyme competitors and the average unrelated distractor are plotted across the first 1600ms post target word onset.

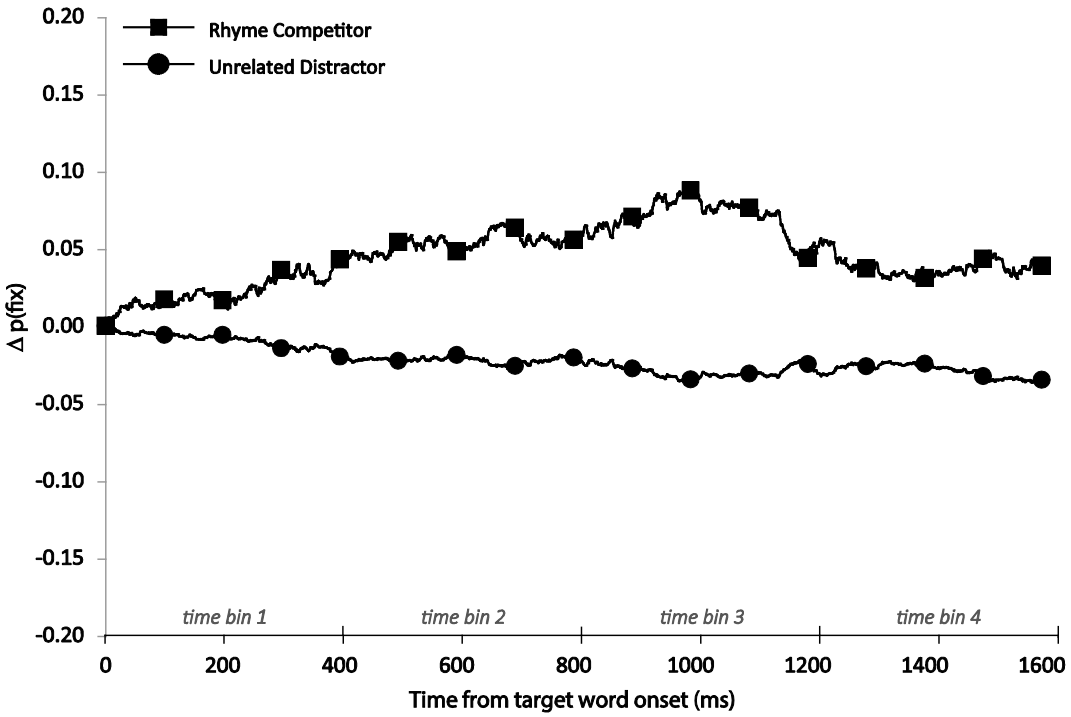

Figure 5: Change in fixation proportions from target word onset in experiment 1. Fixation proportions were averaged across all three unrelated distractors.

To examine the effect of the unfolding spoken target word on fixation behaviour we used a method of analysis similar to that used for analysing simulation results. For analysis we divided the first 1600ms post target word onset into four 400ms bins (1-400ms; 401-800ms; 801-1200ms; 1201-1600ms) and compared behaviour in each of these bins to behaviour in the 400ms that preceded target word onset. For each bin in each trial we calculated the empirical log odds (see Jaeger, 2008) of fixating each category of item (i.e., rhyme competitor, unrelated distractor). The dependent measure was formed by subtracting the log-odds of fixating the unrelated distractor from the log-odds of fixating the phonological rhyme competitor. This difference measure reflects the difference in fixation behaviour as a consequence of phonological overlap. This measure in each of the 400ms time windows post word onset was then compared to the 400ms time window before word onset using linear mixed effect models to examine whether gaze was sensitive to phonological rhyme overlap in each of these post word onset periods. The model used to predict this variable applied the maximal random effect structure (Barr, Levy, Scheepers, & Tily, 2013) with a fixed effect of window and random effects of subject and item. The random effects structure included random intercepts and slopes for time window both by subject and item. To derive p-values we assume t-values were drawn from a normal distribution (Barr, 2008).

A significant effect of phonological rhyme overlap was observed in the second time block (801-1200ms) post word onset [$\beta = 0.68$; $t = 2.22$; $p = 0.03$]. The positive β value indicates that phonological rhyme competitors were fixated above unrelated distractor levels in this time window. A marginal effect of phonological overlap was also observed in the third time block (1201-1600ms) post word onset [$\beta = 0.05$; $t = 1.85$; $p = 0.06$]. There were no statistically robust effects of phonological rhyme overlap in any other time windows.

Discussion

The results of Experiment 1 demonstrated that systematic relationships embedded within the materials, specifically overlap in phonological rhyme shared between spoken target words and visually displayed phonological rhyme competitors, were sufficient to generate a phonological rhyme effect as has previously been described in visual world studies (Allopenna et al., 1998; Huettig & McQueen, 2012; McQueen & Viebahn, 2007), and similar in time course to that observed in Simulation 1.

Experiment 2: Comparing phonological rhyme, visual and semantic overlap effects on language mediated visual attention

To test predictions of the MIM model regarding the relative influence and timing of visual, semantic and phonological rhyme overlap effects on language mediated eye gaze participants gaze was recorded when viewing scenes containing a visual, a semantic and a phonological rhyme competitor in addition to a single unrelated object.

Participants

39 participants (mean age = 25.3, range 18 – 30 years) took part in this study. All were recruited from the MPI for Psycholinguistics subject database and were paid for their participation. All participants were native Dutch speakers and had no known hearing problems and had normal or corrected to normal vision.

Materials

Experiment 2 used the same materials as used in experiment 1 yet with two of the distractors in experimental displays of experiment 1 replaced by a visual competitor and a semantic competitor. Experiment 2 therefore also consisted of 15 experimental trials, 15 target absent filler trials and 15 target present filler trials. On experimental trials, scenes in Experiment 2 therefore now contained a phonological rhyme competitor, a visual competitor, a semantic competitor and an unrelated distractor (see Figure 6). All images of visual and semantic

competitors were black and white line drawings and resized to fit the area 96 x 96 pixels. The four images were arranged evenly in four corners of the display using the same coordinates to centre objects as used in experiment 1. Spoken sentences were also the same as those used in experiment 1. Visual and semantic competitors were monosyllabic words and selected on the basis that they shared a visual or a semantic relationship with the spoken target word. Separate visual (n = 13) and semantic (n = 10) similarity norming studies were conducted to ensure only visual competitors differed from distractors in their level of visual similarity while only semantic competitors differed from distractors in levels of semantic similarity. Semantic competitors were rated marginally more visually similar to target words than unrelated distractors [$\mu = 0.77$; $\sigma = 1.41$ $p = 0.05$], while rhyme competitors were rated as marginally less semantically similar to target words than unrelated distractors [$\mu = -0.41$; $\sigma = 0.85$; $p = 0.08$]. It is likely that participants found it difficult to isolate the effects of visual or semantic similarity from overlap in other dimensions given that rhyme competitors were rated less semantically similar in experiment 2 than in experiment 1, even though participants were required to rate the same rhyme target combinations. Similarity ratings were collected from Dutch native speakers who did not participate in either eye tracking experiment using the same procedure outlined for norming of materials in experiment 1. Further, phonological rhyme competitors were the only set to share an increased level of phonological overlap with the spoken target word. Competitor and distractor sets were also controlled for word frequency, number of letters and number of syllables (see table 3).

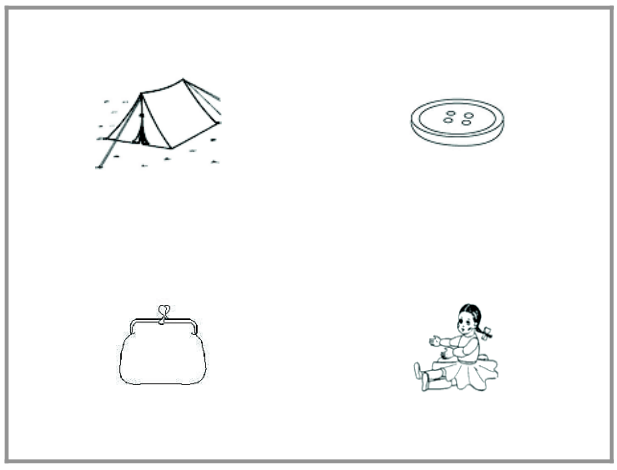

Figure 6: Example of experimental display from experiment 2. Within this trial the target word was "cent", the rhyme competitor was 'tent', the visual competitor 'knoop' (button), the semantic competitor 'beurs' (purse) and the unrelated distractor 'pop' (doll).

Procedure

Experiment 2 followed a procedure identical to that described for experiment 1.

Results

Figure 7 displays the change in the proportion of fixations from target word onset directed towards each category of object (phonological rhyme competitor, visual competitor, semantic competitor, unrelated distractor) in experiment 2 displays for the first 1600ms post word onset.

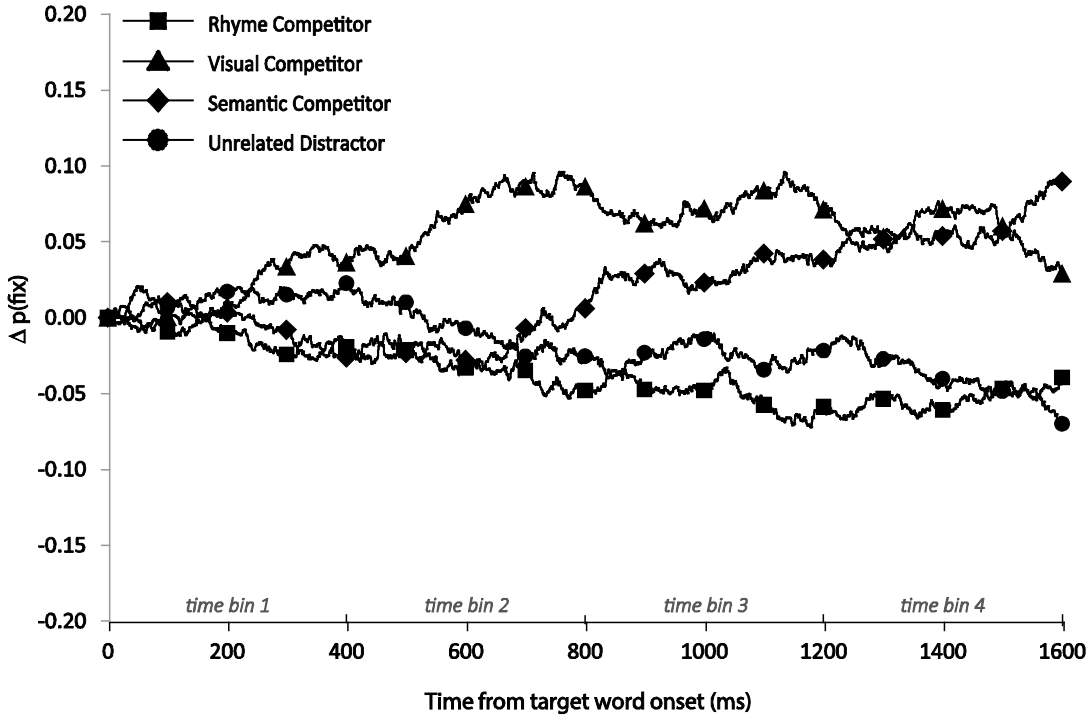

Figure 7: Change in fixation proportions from target word onset directed towards rhyme competitors, visual competitors, semantic competitors and unrelated distractors in experiment 2.

Results of Experiment 2 were analysed using a similar method to that outlined for Experiment 1. However, in Experiment 2 there was a single distractor and three competitors. We therefore compared for each category of competitor (visual, semantic, rhyme) the difference between the empirical log odds of fixating a given competitor and the empirical log odds of fixating the distractor in the 400ms prior to target word onset to the same measure calculated across one of four 400ms time bins post word onset (1-400ms; 401-800ms; 801-1200ms; 1201-1600ms). This analysis revealed that visual competitors were fixated more

than distractors in the second time block (401-800ms) [β = 0.67; t = 2.24; p = 0.03], third time block [β = 0.80; t = 2.68; p = 0.01] and fourth time block [β = 0.58; t = 2.01; p = 0.05]. Semantic competitors were also fixated more than distractors although this effect emerged later, being marginally greater in the third time block [β = 0.45; t = 1.79; p = 0.07] and statistically robust in the fourth time block [β = 0.66; t = 2.51; p = 0.01]. There was, however, no evidence for an influence of phonological rhyme overlap on fixation behaviour as fixation of phonological rhyme competitors did not differ from distractors at any stage post word onset. A post hoc test was also conducted to examine whether the difference between the empirical log odds of fixating the shape competitor and the empirical log odds of fixating the semantic competitor in the 400ms prior to word onset differed from the same difference measure calculated across the entire 1600ms window post word onset, however no significant difference was found [β = 0.3965; t = 1.18; p = 0.24].

Discussion

Results of Experiment 2 show that visual properties shared between the spoken word and visually displayed items are first to bias attention followed by shared semantic properties. Semantic and visual similarity ratings suggest similar levels of overlap exist in the materials between competitor and distractor in both semantic and visual dimensions. This suggests that the initial bias towards visual distractors is driven by underlying architectural constraints of the system driving fixation behaviour or arises due to biases imposed by task specific constraints. For example one explanation may be that visual information is prioritized when processing spoken words under the conditions imposed in this experiment as the task requires a mapping from a spoken word to an item's visual properties. This issue will be discussed further in the context of earlier simulation results in the general discussion section of this paper. Although visual competitors were initially fixated more than semantic competitors, post-hoc analysis shows a similar level of visual and semantic competitor fixation bias across the entire 1600ms post word onset. This indicates that visual and semantic overlap exerts a similar level of influence on language mediated eye gaze. This is similar to a finding reported in Huettig and McQueen (2007) in which visual and semantic competitors were also presented to participants within the same scene. The pattern of results also suggests that the level of overlap implemented in the case of visual and semantic competitors was equivalent in the size of the elicited effect.

In contrast, however, although rhyme competitors only differed from target items in their initial phoneme and therefore overlapped significantly in a phonological dimension with the spoken target word, rhyme competitors failed to attract attention above control levels when visual and semantic competitors were also present. Previous visual world studies (Allopenna et al., 1998; Huettig & McQueen, 2012; McQueen & Viebahn, 2007), including Experiment 1 reported in this paper, have demonstrated that visually displayed items that share their phonological rhyme with a spoken target word do bias fixation behaviour under conditions in which only a systematic phonological relationship exists between displayed items and spoken words. The results of Experiment 2, however, demonstrate that although the level of phonological rhyme overlap embedded in the materials is sufficient to influence eye gaze when phonological rhyme offers the only means of mapping between visually displayed items and spoken words (Experiment 1), this information does not exert an influence when semantic and visual information is also available to map between input streams. These data therefore show that visual and semantic relationships exert a greater influence on language mediated visual attention than phonological rhyme relationships to the extent that even when only a single phoneme in the phonological code mismatches there is no observable influence of phonological rhyme overlap on fixation behaviour. It should be noted, given recent work demonstrating the modulation of phonological rhyme influence by the level of noise in the speech signal (McQueen & Huettig, 2012), that these observed relative influences are likely to vary as a function of environmental factors such as quality of input signals. We debate this point further in the General Discussion.

Irrespective of the relative salience of phonological rhyme information in other conditions the combined data from Experiments 1 and 2 demonstrate the rapid activation of visual and semantic properties when processing spoken words. We know that the level of phonological rhyme overlap embedded in the materials is sufficient to generate an influence on fixation behaviour (experiment 1). Therefore, for there to be no evidence for this effect in Experiment 2 the visual and semantic properties of the spoken target word must have been activated and available to map onto information activated by the visual display before overlapping phonological information in the rhyme of the word could begin to exert an influence on fixation behaviour.

Explicit awareness questionnaire

To assess participants' explicit awareness of the experimental manipulations within each experiment a short questionnaire was completed by participants once they had participated in either of the visual world experiments.

Participants

All participants in experiment 1 (n = 40) and 2 (n = 39) completed the following questionnaire.

Materials & Procedure

Participants were asked to record on paper their response to the following questions: Heb je enige regelmaat kunnen ontdekken in de gerepresenteerde items? (English translation: Did you notice any relationships in the items presented?) to which they could respond 'Ja' (yes) or 'Nee' (No). If they responded 'Ja' then they were requested to provide a written description of the relationships they had noticed (Dutch Instruction: Zo ja, geef een beschrijving).

Results and Discussion

Table 4: Results of experimental manipulation awareness questionnaire.

	Exp. 1 ***(n = 40)***		**Exp. 2** ***(n = 39)***	
	Yes	**No**	**Yes**	**No**
Express awareness	0.45	0.55	0.51	0.49
Identify Rhyme Competitors	0.05	0.95	0.08	0.92
Identify Semantic Competitors	0.03	0.98	0.21	0.79
Identify Visual Competitors	0.03	0.98	0.08	0.92

When scenes contained only rhyme competitors and unrelated distractors 21 of 40 participants indicated that they were not aware of any relationships between the items presented in the experiment. Of the 19 that indicated that they were aware of relationships between items only 2 participants explicitly recorded an awareness of a relationship between the sound of the words presented and items in the display. However, 1 participant recorded an awareness of items sharing a visual relationship, while another participant recorded an

awareness of items sharing a relationship in their meaning even though neither semantic nor visual competitors were present.

In Experiment 2, when displays contained visual competitors, semantic competitors, phonological rhyme competitors and unrelated distractors 19 of 39 participants indicated that they were not aware of any relationships between the items presented in the experiment. Of the 19 that did indicate awareness, 3 indicated an awareness of a relationships between the sound of the word presented and items in the display. 3 participants also indicated an awareness of a visual similarity between items presented. While 8 participants recorded an explicit awareness of a relationship in the meaning of the items presented.

The results of the questionnaire indicate that participants are largely unaware of the experimental manipulations within the materials. Although participants' gaze in Experiment 1 was sensitive to the overlap between the phonological rhyme of the spoken word and that corresponding to the phonological rhyme competitor, only 2 of 40 participants were able to indicate an explicit awareness of this sound similarity. Further, in experiment 2 although robust visual and semantic competitor effects were observed, the vast majority of participants did not register an explicit awareness of similarities between the objects they viewed and words they heard in either of these modalities. The same number of individuals registered an awareness of visual similarity and sound similarity even though in Experiment 2 visual competitor effects were dominant and there was no evidence for sound similarity influencing fixations. Taken together this suggests that the effects observed in both Experiments 1 and 2 represent early implicit processing of the concurrent visual and auditory stimuli that is likely to occur independent of participants' explicit goals, and therefore do not represent strategic processes explicitly engaged by participants given constraints imposed by the experimental manipulation.

4. General Discussion

The purpose of the present study was to examine the interaction of phonological rhyme, visual, and semantic information on language-mediated visual attention. We used the Multimodal Integration Model (MIM) to generate predictions about how eye gaze is distributed across phonological rhyme, visual, and semantic competitors when a system processes in parallel these types of information during spoken word processing. In two visual world eye-tracking experiments we then tested the predictions of the model.

Our experimental results show that shared visual and semantic properties exerted a greater influence on language mediated visual attention than shared properties around the phonological rhyme of words. When visually and semantically related objects were presented in the same scene as objects that share all but their initial phoneme with a spoken target word there was no observable influence of this phonological relationship on fixation behaviour.

Our simulations demonstrate that these findings are compatible with the predictions of a model of language processing in which concurrent phonological, visual and semantic information are integrated in parallel. This work extends the compatibility of the Multimodal Integration Model (MIM) beyond replication of existing word level effects within the visual world literature, to generating hypotheses and predicting behavioural patterns in novel data sets.

As a computational investigation, the Multimodal Integration Model (MIM) provides an explicit description of the connection between underlying cognitive processes and the distribution of eye gaze under visual world conditions. Within the model, relationships between target and competitor are controlled across modalities (a property that is difficult to manipulate with confidence and precision in behavioural studies as our visual and semantic similarity norming studies demonstrate), therefore distinctions in fixation behaviour towards each category of competitor can be isolated to either structural properties of the representations or properties of the system's architecture.

The model predicts the observed pattern of earlier looks towards visual competitors compared to semantic competitors. Within the model this results from semantic competitor effects being more reliant on indirect relationships between stimuli than the visual similarity effects. For semantic competitor effects to emerge, time is required for the target word's phonological input to activate associated semantic properties shared between the semantic competitor and target. Once these shared semantic properties are activated, which are also associated with the visual features of the semantic competitor, fixations towards semantic competitors increase. In contrast, phonological features of the target word are directly associated with the shared visual features of the visual competitor and target, therefore they can begin to influence fixation behaviour as soon as both the visual and phonological stimuli become available.

Within the framework of the model the observed earlier visual effect becomes a property of the experimental design rather than offering evidence that earlier activation of visual information over semantic information is a stable structural property of spoken word

processing. Combining a multimodal computational model with visual world data ensures such errors in inference are avoided. Within the structure of the model, phonological input instead activates associated visual and semantic knowledge in parallel, however as fixations are dependent on the activation of information associated with visual properties of the objects in the immediate visual environment there is a delay before associated shared semantic properties begin to exert an influence on fixations. Pre-activation of semantic properties associated with the semantic competitor would, within this framework, lead to earlier semantic effects, for example in the case that the context prior to word onset induces pre-activation of semantic properties shared between target and semantic competitor. However, as is often the case in visual would studies, should the experimental design ensure that there is no bias in fixations at word onset, then as the properties associated with the visual properties of the semantic competitor (semantic properties) must be activated indirectly, under such conditions the model predicts that looks toward semantic competitors should not precede those toward visual competitors, as these can rely on the direct associations between properties of both stimuli (phonological and visual).

A discrepancy between model predictions and experimental results existed at one point, in that the influence of visual effects in the model exceeded semantic effects at all stages of the test trials. In contrast, although visual effects exceeded semantic effects at early stages of experimental trials, semantic effects exceed visual effects at later stages of the trial in the behavioural data. Although, as previously explained, it would not be possible for the semantic effects to initially exceed visual effects due to the architecture of the model, their relative levels at later stages of the trial is dependent on the strength of semantic associations embedded within the model, a property that is defined by the learning environment. Current simulations show that direct associations between visual and phonological properties are stronger than indirect associations via semantics. This is due to such direct associations being easier to learn. This generates the prediction that at earlier stages of development, visual effects should exceed semantic effects. However, increasing training of the model on both phonological to semantic mappings and semantic knowledge-driven orientation of the word processing system would increase the influence of indirect semantic associations on fixation behaviour, and thus increase the strength of these effects at later stages of test trials once semantic properties have been activated by the phonological input. Increasing training on these tasks requires evidence to show, first, that individuals map to a greater extent from the sound of a word to its meaning than to its visual appearance, and second that individuals

select items based on their semantic (functional) properties more than their phonological properties.

The MIM model also predicted earlier and greater effects of semantic and visual overlap compared to phonological rhyme overlap. This is predicted by the model due to the temporal structure of phonological representations and the continuous parallel multimodal architecture of the model. The nature of processing within the MIM model means that semantic and visual properties associated with initial phonemes are able to exert an effect before visual and semantic properties associated with phonemes in the rhyme. Visual and semantic properties of the visual and semantic competitors are equally likely to be associated with phonemes across the entire target word, therefore they are able to influence fixation behaviour as soon as the target word begins to unfold. In contrast, as only later phonemes are associated with the visual properties of the phonological rhyme competitor, it is only at later stages of the trial that such associations can affect fixation behaviour, hence phonological rhyme effects are delayed in comparison.

Although the level of overlap within the model is controlled across modalities, the model predicts a weaker effect of phonological rhyme overlap compared to visual and semantic overlap. Again this is a product of the temporal structure of phonological representations and the parallel architecture of the model. The architecture facilitates associations between visual, semantic and phonological properties at all stages of processing, therefore as a spoken word unfolds visual and semantic information associated with initial phonological information can begin to constrain processing. As previously stated this means that in the case of visual and semantic competitors associations between the unfolding word and its visual and semantic properties are able to develop and thus influence fixation behaviour during test trials at all stages of phonological processing. This also means that by the time information carried in the phonological rhyme becomes available there is often sufficient information with which to identify the target. For this reason visual and semantic associations with phonological properties of the rhyme are weaker and slower to develop, hence phonological rhyme competitor effects that are dependent on such associations are weaker.

Although the model correctly predicted greater visual and semantic effects than phonological rhyme effects, activation of eye layer units corresponding to the phonological rhyme competitor still exceeded those of units corresponding to unrelated distractors. Hence, should eye layer activity relate directly to the probability of fixating an object, the model would

predict phonological rhyme effects in Experiment 2. We do not believe this to compromise the model's validity, however, as saccades in the model could be restricted to be only initiated to a location when activation of the corresponding eye layer unit exceeds pre-word onset levels. In Experiment 1, eye layer units corresponding to the phonological rhyme competitor increase from baseline once the spoken word begins to unfold and remain above pre-word onset levels throughout the trial. In contrast, in Experiment 2, eye layer units corresponding to the phonological rhyme competitor decrease in activation relative to their level of activation prior to word onset, therefore such a mechanism would generate the observed pattern of results, a phonological rhyme effect in Experiment 1 yet not in Experiment 2. In this case, the apparent influence of the phonological rhyme competitors on processing late in the trial for the model is taken to indicate low-level, but below-threshold, activation of the rhyme competitor in the behavioural data, thus, the model's processing can provide even greater sensitivity to the relative contributions of information streams as word processing unfolds, than are possible from behavioural investigations.

A critical feature of the model, that influences the structure of effects observed, is the level of noise to which the model is exposed in the learning environment. The strength of associations developed between representations in visual, semantic and phonological dimensions is dependent on the reliability of the signal in each modality. For example, associations between phonological properties in the rhyme of a word and the word's corresponding visual and semantic properties increase once noise is introduced to the speech signal such that phonological properties in the onset are no longer perfect predictors of the target (Smith, Monaghan & Huettig, 2013). Within the model the same mechanism dictates the strength of associations across modalities, therefore manipulating levels of noise in visual and semantic dimensions is likely to also modulate the visual and semantic effects observed. This predicts that populations that differ in their levels of exposure to noise in one of these modalities, e.g., visually or hearing impaired, will display distinct patterns of sensitivity to visual, semantic and phonological overlap in the visual world paradigm. McQueen & Huettig, (2012) demonstrate that the influence of phonological rhyme information can also be dynamically adjusted by short term exposure to noise in the speech signal in a manner similar to that captured by the model when noise levels are manipulated in training. It remains to be seen whether noise in other modalities can dynamically adjust sensitivity to specific visual or semantic properties in the visual world paradigm as the model would predict.

The data from Experiments 1 and 2 demonstrate that visual and semantic information is activated rapidly by the incoming speech signal and can be recruited by the cognitive system to map onto pre-activated information (e.g. activated via the immediate visual environment). The speed with which such multimodal activation and integration occurs in Experiment 2 is sufficient to ensure that information activated by later phonemes in the rhyme of a word exert no observable influence on behaviour, even though Experiment 1 demonstrates that sufficiently strong relationships exist within the materials to generate detectable effects. This offers support for models of spoken word processing in which multimodal information is activated and integrated early via either a rapid cascading or parallel process. This argument is further supported by the compatibility of such observations with predictions generated by the MIM model of spoken word processing in which concurrent visual, semantic and phonological information is integrated in parallel. In which many of the subtle features of behaviour captured by the two visual world studies are as we have argued emergent properties of the parallel architecture implemented in the model.

We acknowledge that language mediated visual attention only offers an indirect measure of the processes underlying spoken word recognition and comprehension. Our data nevertheless clearly show that visual and semantic information is activated rapidly and available to constrain behaviour as a spoken word unfolds. Results of the post-experiment questionnaire also suggest that effects captured by the visual world experiments in this paper reflect early implicit processing that is likely to occur independent of a participant's explicit awareness or goals. Given the ambiguities present in natural language (Ferreira, 2008; Jaeger, 2006, 2010; Piantadosi, Tily & Gibson, 2012; Roland, Elman & Ferreira, 2006; Wasow & Arnold, 2003; Wasow et al., 2005) and given, as our results demonstrate, that the cognitive system has access to multimodal information in which reliable cues as to the meaning of a given utterance are likely to exist, an efficient spoken word recognition system should rapidly accommodate such cues adapting its response in accordance to the current multimodal evidential landscape. As the ubiquity of ambiguity within natural language is rarely harmful to communication, it has been assumed that integration of contextual information is "cognitively cheap" (Levinson, 2000). Within this paper we present a multimodal constraint satisfaction model that offers an explicit description of an architecture able to support such efficient processing.

Our experimental data show that under conditions in which visual, semantic and phonological rhyme information are all available to constrain word referent mapping, visual and semantic

relationships dominate such that phonological rhyme exerts no observable influence on behaviour. We argue that although noise is likely to influence the relative contribution of information streams, that during every-day spoken word processing, in situations where visual and semantic information is likely to be pre-activated, phonological rhyme information may exert little or no *observable* influence on processing.

To summarize, models of speech recognition have frequently overlooked the multimodal nature of the speech recognition problem in "real world" environments. Most past studies have focussed purely on the phonological properties of the system (e.g. Luce et al., 2000; McClelland & Elman, 1986; Norris & McQueen, 2008; Scharenborg & Boves, 2010). Our data strongly suggest that this is likely to describe only a single component of a complex multimodal system. We therefore conclude that for a comprehensive description of both spoken word processing, multimodal models, such as the multimodal integration model (MIM), tested here, are required.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419-439.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247-264.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163-187.

Barr, D. J. (2008). Analyzing 'visual world'eyetracking data using multilevel logistic regression. *Journal of memory and language*, *59*(4), 457-474.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition?. *Journal of Memory and Language*, *32*(2), 193-210.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*, 84−107.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 498-513.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review*, *12*(3), 453-459.

Duñabeitia, J. A., Avilés, A., Afonso, O., Scheepers, C., & Carreiras, M. (2009). Qualitative differences in the representation of abstract versus concrete words: Evidence from the visual-world paradigm. *Cognition*, *110*(2), 284-292.

Ferreira, V. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation*, *49*, 209–246.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT press.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*(5-6), 613-656.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, *45*(2), 220–266.

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), B23-B32.

Huettig, F., & Altmann, G. T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, *15*(8), 985-1018.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*(4), 460-482.

Huettig, F., & Altmann, G. T. (2011). Looking at anything that is green when hearing “frog”: How object surface colour and stored object colour knowledge influence language-mediated overt attention. *The Quarterly Journal of Experimental Psychology*, *64*(1), 122-145.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151-171.

Jaeger, T. (2006). *Redundancy and syntactic reduction in spontaneous speech. Unpublished doctoral dissertation*, Stanford University.

Jaeger, T. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62.

Kukona, A., & Tabor, W. (2011). Impulse processing: A dynamical systems model of incremental eye movements in the visual world paradigm. *Cognitive Science*, *35*(6), 1009-1051.

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics*, *62*(3), 615-625.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4), 676.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: studies with artificial lexicons. *Journal of Experimental Psychology: General*, *132*(2), 202-227.

Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2000). Simple recurrent networks and competition effects in spoken word recognition. *University of Rochester Working Papers in Language Science*, *1*, 56-71.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71-102.

Mayberry, M. R., Crocker, M. W., & Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science*, *33*(3), 449-496.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1-86.

McClelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science*, *38*(6), 1139-1189.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The Appeal of Parallel Distributed Processing. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition. Volume I. (pp. 3–44). Cambridge, MA: MIT Press.

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology*, *60*(1), 1-39.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America*, *131*(1), 509-517.

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *The Quarterly Journal of Experimental Psychology*, *60*(5), 661-671.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & Cognition*, *37*(7), 1026-1039.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475-494.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189-234.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, 1(2), 263–269.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280-291.

Pulvermüller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language*, *110*(2), 81-94.

R Development Core Team. (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science*, *38*(6), 1024-1077.

Roland, D., Elman, J. L., & Ferreira, V. S. (2006). Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*, *98*(3), 245-272.

Rumelhart, D. E., & McClelland, J. L., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: Foundations & volume II: Psychological and biological models.* Cambridge, MA: MIT Press.

Scharenborg, O., & Boves, L. (2010). Computational modelling of spoken-word recognition processes: Design choices and evaluation. *Pragmatics & Cognition*, *18*(1), 136-164.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Smith, A., Monaghan, P., & Huettig, F. (2013). An amodal shared resource model of language-mediated visual attention. *Frontiers in Psychology*, 4: 528.

Smith, A., Monaghan, P., & Huettig, F. (2014). Modelling language – vision interactions in the hub and spoke framework. In J. Mayor, & P. Gomez (Eds.), *Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop (NCPW13)*. (pp. 3-16). Singapore: World Scientific Publishing.

Spivey, M. (2007). *The continuity of mind*. Oxford: Oxford University Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632-1634.

Wasow, T., & Arnold, J. (2003). Post-verbal constituent ordering in English. *Determinants of Grammatical Variation in English*, 119–154.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe.* Stanford: CSLI Publications.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 1-14.

Appendix

Neural networks simulations were conducted using Mikenet version 8.0 developed by M. W. Harm (www.cnbc.cmu.edu/~mharm/research/tools/mikenet/), a collection of libraries written in the C programming language for implementing and training connectionist networks.

Networks were trained using the continuous recurrent backpropagation through time training algorithm provided in Mikenet (crbp.c) which implements Pearlmutter (1989). Unit activation was calculated using a logistic activation function and sum squared error was used to calculate error. Time within the network was modelled using 14 samples and an integration constant of 0.25. All other parameters were set to the default values implemented in Mikenet version 8.0.

Mixed effects model analysis was performed using the R (version 3.1.0; R Development Core Team, 2009) libraries lme4 (version 1.1-6) and languageR (version 1.4.1).

Table A.1: Simulation 1 t-test results for time steps post word onset

By Item							By Instantiation						
Competitor	Time Step	Ratio	Confidence Interval		t-value	p-value	Competitor	Time Step	Ratio	Confidence Interval		t-value	p-value
			Lower	Upper						Lower	Upper		
Phonological	1	0.000	0.000	0.000	NA	NA	Phonological	1	0.000	0.000	0.000	NA	NA
	2	-0.022	-0.053	0.009	-1.479	0.155		2	-0.017	-0.064	0.030	-0.859	0.419
	3	-0.023	-0.062	0.017	-1.199	0.245		3	-0.017	-0.057	0.023	-1.010	0.346
	4	-0.027	-0.079	0.024	-1.122	0.276		4	-0.020	-0.065	0.025	-1.048	0.330
	5	-0.035	-0.081	0.011	-1.596	0.127		5	-0.026	-0.061	0.008	-1.794	0.116
	6	-0.020	-0.086	0.046	-0.626	0.539		6	-0.012	-0.062	0.038	-0.564	0.591
	7	0.063	-0.038	0.164	1.306	0.207		7	0.068	-0.013	0.148	1.994	0.086
	8	0.157	-0.022	0.336	1.835	0.082		8	0.159	0.005	0.312	2.439	0.045
	9	0.323	0.073	0.574	2.705	0.014		9	0.309	0.106	0.512	3.605	0.009
	10	0.471	0.145	0.797	3.023	0.007		10	0.434	0.206	0.662	4.507	0.003
	11	0.555	0.196	0.913	3.239	0.004		11	0.511	0.303	0.719	5.813	0.001
	12	0.594	0.212	0.976	3.256	0.004		12	0.546	0.302	0.790	5.289	0.001
	13	0.575	0.197	0.953	3.181	0.005		13	0.535	0.307	0.762	5.548	0.001
	14	0.554	0.178	0.931	3.084	0.006		14	0.517	0.280	0.753	5.173	0.001
	15	0.552	0.202	0.903	3.296	0.004		15	0.531	0.252	0.809	4.505	0.003
	16	0.546	0.205	0.887	3.353	0.003		16	0.533	0.264	0.803	4.677	0.002
	17	0.484	0.171	0.797	3.234	0.004		17	0.472	0.245	0.699	4.922	0.002
	18	0.442	0.131	0.753	2.978	0.008		18	0.427	0.204	0.650	4.522	0.003
	19	0.408	0.090	0.726	2.682	0.015		19	0.394	0.158	0.630	3.944	0.006
	20	0.420	0.093	0.747	2.690	0.015		20	0.407	0.153	0.662	3.787	0.007
	21	0.427	0.089	0.765	2.645	0.016		21	0.405	0.181	0.630	4.264	0.004
	22	0.446	0.107	0.785	2.752	0.013		22	0.422	0.193	0.650	4.365	0.003
	23	0.471	0.107	0.835	2.710	0.014		23	0.439	0.226	0.652	4.879	0.002
	24	0.494	0.129	0.858	2.833	0.011		24	0.464	0.257	0.671	5.293	0.001
	25	0.534	0.176	0.892	3.124	0.006		25	0.508	0.260	0.755	4.848	0.002

Table A.2: Simulation 2 t-test results for time steps post word onset

By Item							By Instantiation						
Competitor	Time Step	Ratio	Confidence Interval		t-value	p-value	Competitor	Time Step	Ratio	Confidence Interval		t-value	p-value
			Lower	Upper						Lower	Upper		
Phonological	1	0.000	0.000	0.000	NA	NA	Phonological	1	0.000	0.000	0.000	NA	NA
	2	0.042	-0.017	0.101	1.494	0.152		2	0.007	-0.050	0.064	0.304	0.770
	3	0.110	-0.010	0.229	1.926	0.069		3	0.039	-0.072	0.149	0.831	0.433
	4	0.112	-0.025	0.250	1.707	0.104		4	0.044	-0.113	0.200	0.659	0.531
	5	0.152	0.000	0.304	2.093	0.050		5	0.082	-0.123	0.286	0.943	0.377
	6	0.163	0.004	0.321	2.149	0.045		6	0.127	-0.153	0.407	1.071	0.320
	7	0.117	-0.103	0.337	1.116	0.278		7	0.064	-0.271	0.398	0.450	0.666
	8	0.151	-0.098	0.400	1.269	0.220		8	0.188	-0.321	0.696	0.873	0.411
	9	0.246	-0.150	0.642	1.299	0.210		9	0.272	-0.257	0.802	1.217	0.263
	10	0.347	-0.248	0.941	1.221	0.237		10	0.270	-0.123	0.663	1.625	0.148
	11	0.422	-0.229	1.072	1.358	0.191		11	0.307	-0.108	0.723	1.748	0.124
	12	0.346	-0.122	0.815	1.547	0.139		12	0.380	-0.142	0.903	1.721	0.129
	13	0.492	-0.038	1.022	1.945	0.067		13	0.471	0.043	0.899	2.600	0.035
	14	0.526	-0.016	1.068	2.032	0.056		14	0.509	0.074	0.944	2.767	0.028
	15	0.525	-0.010	1.061	2.055	0.054		15	0.498	0.063	0.934	2.708	0.030
	16	0.486	-0.011	0.983	2.046	0.055		16	0.461	0.111	0.811	3.116	0.017
	17	0.483	0.100	0.866	2.640	0.016		17	0.495	0.097	0.893	2.942	0.022
	18	0.621	0.193	1.050	3.036	0.007		18	0.539	0.118	0.961	3.024	0.019
	19	0.707	0.101	1.313	2.440	0.025		19	0.532	0.087	0.976	2.829	0.026
	20	0.687	0.005	1.368	2.110	0.048		20	0.552	0.024	1.080	2.471	0.043
	21	1.072	-0.302	2.446	1.633	0.119		21	0.623	-0.095	1.341	2.051	0.080
	22	0.852	0.040	1.664	2.195	0.041		22	0.670	0.036	1.304	2.500	0.041
	23	1.013	0.055	1.972	2.213	0.039		23	0.694	0.195	1.192	3.291	0.013
	24	1.020	0.034	2.005	2.166	0.043		24	0.677	0.225	1.129	3.540	0.010
	25	0.792	0.037	1.548	2.194	0.041		25	0.616	0.260	0.972	4.089	0.005
Visual	1	0.000	0.000	0.000	NA	NA	Visual	1	0.000	0.000	0.000	0.000	NA
	2	0.014	-0.034	0.062	0.608	0.550		2	-0.009	-0.071	0.053	-0.347	0.739
	3	0.057	-0.028	0.142	1.407	0.176		3	0.007	-0.067	0.080	0.214	0.836
	4	0.055	-0.049	0.160	1.107	0.282		4	0.009	-0.081	0.100	0.239	0.818
	5	0.111	-0.003	0.225	2.041	0.055		5	0.056	-0.065	0.178	1.096	0.309
	6	0.166	0.007	0.324	2.183	0.042		6	0.155	-0.053	0.362	1.762	0.121
	7	0.258	0.006	0.510	2.139	0.046		7	0.192	-0.024	0.408	2.100	0.074
	8	0.361	0.174	0.547	4.046	0.001		8	0.388	0.040	0.736	2.633	0.034
	9	0.580	0.249	0.911	3.668	0.002		9	0.615	0.134	1.097	3.023	0.019
	10	0.828	0.387	1.268	3.934	0.001		10	0.811	0.458	1.165	5.423	0.001
	11	1.214	0.683	1.745	4.787	0.000		11	1.149	0.792	1.507	7.594	0.000
	12	1.536	0.892	2.180	4.990	0.000		12	1.491	0.960	2.022	6.639	0.000
	13	1.928	1.226	2.631	5.744	0.000		13	1.822	1.323	2.321	8.631	0.000
	14	2.199	1.440	2.957	6.067	0.000		14	2.105	1.595	2.615	9.756	0.000
	15	2.150	1.439	2.862	6.325	0.000		15	2.113	1.487	2.739	7.984	0.000
	16	2.220	1.431	3.008	5.892	0.000		16	2.116	1.496	2.736	8.070	0.000
	17	2.539	1.489	3.590	5.059	0.000		17	2.236	1.490	2.982	7.092	0.000
	18	3.081	1.362	4.800	3.752	0.001		18	2.207	1.482	2.932	7.196	0.000
	19	3.083	1.494	4.671	4.062	0.001		19	2.272	1.330	3.214	5.703	0.001
	20	2.775	1.503	4.048	4.564	0.000		20	2.257	1.068	3.446	4.489	0.003
	21	3.350	1.578	5.122	3.956	0.001		21	2.430	0.960	3.900	3.908	0.006
	22	2.941	1.628	4.254	4.689	0.000		22	2.428	1.283	3.572	5.014	0.002
	23	3.084	1.709	4.460	4.693	0.000		23	2.439	1.760	3.118	8.497	0.000
	24	3.174	1.855	4.493	5.035	0.000		24	2.553	1.902	3.204	9.273	0.000
	25	2.791	1.818	3.765	6.000	0.000		25	2.440	1.941	2.939	11.571	0.000
Semantic	1	0.000	0.000	0.000	NA	NA	Semantic	1	0.000	0.000	0.000	NA	NA
	2	0.058	-0.024	0.141	1.475	0.157		2	0.016	-0.024	0.056	0.937	0.380
	3	0.100	-0.013	0.214	1.848	0.080		3	0.029	-0.027	0.084	1.225	0.260
	4	0.100	-0.025	0.225	1.673	0.111		4	0.028	-0.038	0.094	1.003	0.349
	5	0.121	-0.031	0.273	1.672	0.111		5	0.067	-0.034	0.168	1.575	0.159
	6	0.114	-0.057	0.284	1.396	0.179		6	0.095	-0.019	0.209	1.965	0.090
	7	0.149	-0.071	0.368	1.418	0.173		7	0.086	-0.138	0.310	0.906	0.395
	8	0.200	0.016	0.385	2.271	0.035		8	0.234	-0.117	0.586	1.575	0.159
	9	0.374	0.007	0.740	2.134	0.046		9	0.446	-0.014	0.906	2.295	0.055
	10	0.583	-0.007	1.172	2.068	0.053		10	0.587	0.123	1.051	2.989	0.020
	11	1.013	0.132	1.894	2.407	0.026		11	0.855	0.465	1.246	5.178	0.001
	12	1.284	0.433	2.135	3.157	0.005		12	1.178	0.622	1.733	5.012	0.002
	13	1.676	0.691	2.662	3.559	0.002		13	1.430	0.874	1.986	6.080	0.001
	14	1.878	0.941	2.816	4.195	0.001		14	1.684	0.976	2.391	5.628	0.001
	15	1.977	1.022	2.932	4.333	0.000		15	1.774	1.009	2.540	5.483	0.001
	16	1.896	1.074	2.719	4.827	0.000		16	1.775	1.039	2.510	5.706	0.001
	17	2.074	1.259	2.889	5.327	0.000		17	1.880	1.113	2.647	5.795	0.001
	18	2.431	1.241	3.621	4.277	0.000		18	1.842	1.097	2.588	5.844	0.001
	19	2.477	1.166	3.788	3.953	0.001		19	1.816	0.996	2.635	5.239	0.001
	20	2.477	0.927	4.028	3.343	0.003		20	1.842	0.839	2.846	4.342	0.003
	21	3.558	0.075	7.042	2.138	0.046		21	1.996	0.727	3.266	3.720	0.008
	22	2.726	0.984	4.468	3.276	0.004		22	2.046	0.869	3.223	4.112	0.005
	23	2.959	1.056	4.861	3.254	0.004		23	2.065	1.186	2.945	5.552	0.001
	24	3.218	1.191	5.245	3.323	0.004		24	2.204	1.305	3.104	5.794	0.001
	25	2.881	1.397	4.365	4.064	0.001		25	2.258	1.418	3.098	6.355	0.000

Table 3: Mixed effect models analysis results

Simulation	Competitor	Time Window	Estimate	Std. Error	t- value	p-value
1	Rhyme - Distractor	6 - 10	-0.190	0.091	-2.086	0.037
		11 - 15	0.075	0.073	1.031	0.302
		16 - 20	0.367	0.134	2.730	0.006
		21 - 25	0.300	0.121	2.479	0.013
		26 - 30	0.304	0.131	2.324	0.020
2	Rhyme - Distractor	6 - 10	0.018	0.132	0.137	0.891
		11 - 15	0.110	0.128	0.861	0.389
		16 - 20	0.258	0.136	1.896	0.058
		21 - 25	0.300	0.119	2.531	0.011
		26 - 30	0.330	0.110	3.008	0.003
	Visual - Distractor	6 - 10	-0.110	0.119	-0.919	0.358
		11 - 15	0.262	0.102	2.567	0.010
		16 - 20	0.950	0.134	7.074	< 0.001
		21 - 25	1.118	0.143	7.797	< 0.001
		26 - 30	1.158	0.133	8.731	< 0.001
	Semantic - Distractor	6 - 10	0.038	0.134	0.287	0.774
		11 - 15	0.232	0.107	2.164	0.030
		16 - 20	0.782	0.153	5.124	< 0.001
		21 - 25	0.938	0.147	6.379	< 0.001
		26 - 30	0.989	0.175	5.646	< 0.001
	Visual - Semantic	6 - 10	-0.148	0.133	-1.113	0.266
		11 - 15	0.030	0.108	0.278	0.781
		16 - 20	0.168	0.163	1.034	0.301
		21 - 25	0.180	0.163	1.105	0.269
		26 - 30	0.169	0.174	0.974	0.330
1 vs 2	Rhyme: Exp 1 - Exp 2	6 - 10	0.200	0.087	2.284	0.022
		11 - 15	0.113	0.078	1.460	0.144
		16 - 20	0.042	0.067	0.621	0.535
		21 - 25	0.096	0.067	1.423	0.155
		26 - 30	0.109	0.067	1.627	0.104
	Rhyme: Exp*Bin	6 - 10	0.208	0.141	1.477	0.140
		11 - 15	0.035	0.117	0.301	0.764
		16 - 20	-0.108	0.131	-0.828	0.408
		21 - 25	0.000	0.133	0.000	1.000
		26 - 30	0.026	0.134	0.195	0.845

Table A.3: Experiment 1 materials, Rhyme Competitor Only

Spoken Target				Rhyme Competitor							Distractor Set 1							Distractor Set 2							Distractor Set 3						
Word	Freq	Letter	Syll	Word	Freq	Letter	Syll	Pho	Sem†	Vis††	Word	Freq	Letter	Syll	Pho	Sem†	Vis††	Word	Freq	Letter	Syll	Pho	Sem†	Vis††	Word	Freq	Letter	Syll	Pho	Sem†	Vis††
boek	250	4	1	koek	7	4	1	2	1.9	1.9	kennel	1	6	2	0	0.9	0.8	neus	98	4	1	0	1.5	2.3	auto	165	4	2	0	1.6	1.6
bord	38	4	1	fort	14	4	1	3	1.2	1.1	tank	7	4	1	1	2.0	1.8	kist	29	4	1	1	1.4	0.5	trein	70	5	1	1	2.7	0.2
cent	18	4	1	tent	20	4	1	3	0.9	0.5	pop	11	3	1	0	1.1	2.0	ster	14	4	1	3	1.5	0.1	fles	74	4	1	1	2.0	2.2
dolk	5	4	1	wolk	15	4	1	3	1.4	0.1	brug	41	4	1	0	1.3	1.3	spin	6	4	1	0	2.3	2.2	boot	49	4	1	0	2.2	2.8
garen	2	5	2	varen	14	5	2	3	1.6	1.8	eend	12	4	1	0	0.5	2.4	ijsje	3	5	2	1	0.7	1.1	slot	70	4	1	0	1.5	1.2
grill	1	5	1	bril	32	4	1	3	1.9	0.9	kaart	53	5	1	1	1.3	0.1	zwaan	4	5	1	0	0.6	1.6	fluit	1	5	1	1	0.7	1.8
hoed	31	4	1	voet	96	4	1	2	0.5	1.1	kassa	6	5	2	0	1.5	0.8	raam	112	4	1	0	2.1	1.5	zaag	3	4	1	0	0.7	0.4
jurk	34	4	1	kurk	1	4	1	3	1.2	1.9	penseel	4	7	2	0	2.1	2.8	lamp	21	4	1	0	1.6	1.2	kano	3	4	2	1	0.9	1.2
maan	62	4	1	haan	12	4	1	2	1.9	3.4	boom	53	4	1	1	0.3	0.3	vork	10	4	1	0	3.5	2.3	riem	13	4	1	0	0.6	4.3
mand	12	4	1	tand	12	4	1	3	0.6	1.3	brief	114	5	1	0	1.6	1.8	vest	8	4	1	1	2.1	0.9	worm	3	4	1	1	0.5	0.3
muis	9	4	1	buis	1	4	1	2	1.7	1.7	schaar	5	6	1	0	1.6	1.9	tang	3	4	1	0	1.4	1.6	vlag	18	4	1	0	0.8	0.5
paal	9	4	1	sjaal	6	5	1	2	1.5	2.0	slee	2	4	1	1	1.3	0.1	duif	8	4	1	0	2.7	3.3	beker	9	5	2	0	1.9	3.4
pijl	9	4	1	teil	3	4	1	2	1.5	0.4	broek	56	5	1	0	1.3	1.8	teen	7	4	1	0	1.3	3.6	kaas	43	4	1	0	0.9	0.2
riool	2	5	2	viool	9	5	2	4	1.8	1.0	pak	46	3	1	0	1.4	1.7	berg	21	4	1	0	0.8	0.8	arend	3	5	2	0	1.8	0.5
roos	14	4	1	doos	22	4	1	2	0.9	0.6	servet	4	6	2	1	2.2	0.3	beer	14	4	1	1	1.3	1.5	haai	1	4	1	0	2.0	0.8
μ	33.07	4.20	1.13		17.60	4.20	1.13	2.60	1.37	1.31		27.67	4.73	1.27	0.33	1.36	1.33		23.87	4.13	1.07	0.47	1.65	1.65		35.00	4.27	1.27	0.33	1.41	1.42
σ	62.36	0.41	0.35		23.24	0.41	0.35	0.63	0.46	0.84		32.16	1.16	0.46	0.49	0.53	0.88		33.88	0.35	0.26	0.83	0.78	0.97		45.50	0.46	0.46	0.49	0.69	1.26

† n = 11
†† n = 13

Table A.4: Experiment 2 materials, Rhyme, Semantic and Visual Competitors

Spoken Target				Rhyme Competitor							Visual Competitor							Semantic Competitor							Distractor						
Word	Freq	Letter	Syll	Word	Freq	Letter	Syll	Pho	Sem*	Vis**	Word	Freq	Letter	Syll	Pho	Sem*	Vis**	Word	Freq	Letter	Syll	Pho	Sem*	Vis**	Word	Freq	Letter	Syll	Pho	Sem*	Vis**
boek	250	4	1	koek	7	4	1	2	0.8	1.1	deur	325	4	1	0	1.6	6.6	pen	16	3	1	0	5.7	1.8	kennel	1	6	2	0	0.6	1.6
bord	38	4	1	fort	14	4	1	3	1.1	0.6	stuur	25	5	1	1	2.6	6.7	vork	10	4	1	2	7.0	1.2	kat	49	3	1	1	2.3	0.6
cent	18	4	1	tent	20	4	1	3	0.5	0.6	knoop	13	5	1	0	3.6	8.8	beurs	9	5	1	1	8.1	1.8	pop	11	3	1	0	1.5	0.2
dolk	5	4	1	wolk	15	4	1	3	1.0	0.9	veer	2	4	1	0	1.7	6.1	schild	7	6	1	1	5.4	1.1	brug	41	4	1	0	1.5	2.3
garen	2	5	2	varen	14	5	2	3	1.6	2.2	slang	18	5	1	0	1.1	5.7	rits	3	4	1	0	5.8	4.0	eend	12	4	1	0	0.2	0.8
grill	1	5	1	bril	32	4	1	3	1.5	0.8	hek	25	3	1	0	2.5	7.2	lucifer	8	7	3	1	5.8	1.9	kaart	53	5	1	0	1.4	2.2
hoed	31	4	1	voet	96	4	1	2	0.7	1.0	bel	19	3	1	0	2.0	7.5	jas	42	3	1	0	6.5	1.7	kassa	6	5	2	0	1.2	1.7
jurk	34	4	1	kurk	1	4	1	3	0.8	2.6	vaas	7	4	1	0	1.8	5.5	riem	13	4	1	0	6.0	1.2	penseel	4	7	2	0	1.8	1.8
maan	62	4	1	haan	12	4	1	2	1.8	0.4	tomaat	2	6	2	2	0.9	7.5	ster	14	4	1	0	7.6	3.7	boom	53	4	1	0	2.5	1.0
mand	12	4	1	tand	12	4	1	3	0.3	1.4	trom	2	4	1	1	1.8	7.0	fles	74	4	1	0	2.9	2.1	brief	114	5	1	0	2.0	0.7
muis	9	4	1	buis	1	4	1	2	1.0	2.2	jojo	0	4	2	0	0.5	4.7	worm	3	4	1	1	5.0	2.1	schaar	5	6	1	1	0.7	0.8
paal	9	4	1	sjaal	6	5	1	2	0.6	2.1	rietje	2	6	2	0	2.6	7.5	trap	90	4	1	0	3.9	3.9	slee	2	4	1	1	1.7	2.8
pijl	9	4	1	teil	3	4	1	2	1.0	0.4	haai	1	4	1	0	2.7	4.8	ridder	7	6	2	0	5.8	1.7	broek	56	5	1	0	0.9	3.5
riool	2	5	2	viool	9	5	2	4	1.3	1.5	fluit	5	5	1	1	0.8	5.9	toilet	14	6	2	1	7.9	3.5	pak	46	3	1	0	0.6	1.5
roos	14	4	1	doos	22	4	1	2	0.2	0.6	lamp	21	4	1	0	0.6	5.9	parfum	13	6	2	0	5.1	3.0	servet	4	6	2	1	1.5	1.6
μ	33.07	4.20	1.13		17.60	4.20	1.13	2.60	0.95	1.23		31.13	4.40	1.20	0.33	1.78	6.51		21.53	4.67	1.33	0.47	5.90	2.30		30.47	4.67	1.27	0.27	1.36	1.53
σ	62.36	0.41	0.35		23.24	0.41	0.35	0.63	0.46	0.74		81.82	0.91	0.41	0.62	0.89	1.13		26.37	1.23	0.62	0.64	1.42	1.03		32.06	1.23	0.46	0.46	0.67	0.90

* n = 10
** n = 13

Freq = word frequency
Letter = no. letters
Syll = no. syllables
Pho = number of overlapping phonemes with spoken target word
Sem = semantic similarity rating
Vis = visual similarity rating

Chapter 5

Literacy effects on language and vision: Emergent effects from the multimodal integration model (MIM)[1]

Abstract

Learning to read and write requires an individual to connect additional orthographic representations to pre-existing mappings between phonological and semantic representations of words. Past empirical results suggest that the process of learning to read and write (at least in alphabetic languages) elicits changes in the language processing system, by either increasing the cognitive efficiency of mapping between representations associated with a word, or by changing the granularity of phonological processing of spoken language, or through a combination of both. Behavioural effects of literacy have typically been assessed in offline explicit tasks that have addressed only phonological processing. However, a recent eye tracking study compared high and low literate participants on effects of phonology and semantics in processing measured implicitly using eye movements. High literates' eye movements were more affected by phonological overlap in online speech than low literates, with only subtle differences observed in semantics. We determined whether these effects were due to cognitive efficiency and/or granularity of speech processing in a multimodal model of speech processing – the multimodal integration model (MIM, Smith, Monaghan, & Huettig, 2013). We found that cognitive efficiency in the model had only a marginal effect on semantic processing and did not affect performance for phonological processing, whereas fine-grained versus coarse-grained phonological representations in the model simulated the high/low literacy effects on phonological processing, suggesting that literacy has a focused effect in changing the grain-size of phonological mappings.

[1] *Adapted from: Smith, A. C., Monaghan, P., & Huettig, F. (2014). Literacy effects on language and vision: Emergent effects from an amodal shared resource (ASR) computational model. Cognitive Psychology. 75, 28-54.*

1. Introduction

Approximately 16% of the world's adult population are illiterate, defined as "the ability to read and write with understanding a simple statement related to one's daily life" (UNESCO Institute for Statistics, 2013). Learning to read has a profound effect on cognitive processing, resulting in qualitative changes to the representation of phonological information about words, but also correlating with a general increase in cognitive processing performance. Much of our understanding of language processing is based on data and theoretical and computational models only of literate participants, but a full understanding of language comprehension and production must also take into account the role of literacy in processing. Previous models of literacy effects on language processing have not effectively distinguished between accounts based on a general cognitive ability increase and more specific phonological processing changes.

Here, we test an implemented computational model of language processing that was previously applied only to data from literate participants. We extended the model to simulate both the general cognitive processing account as well as the phonological representation account in order to account for data from literate and illiterate participants in language processing tasks. We first review the two theoretical accounts of effects of literacy on language processing – the phonological processing change and the general cognitive processing accounts – before describing previous models of effects of literacy on language processing. We then present the advantages of a language processing task that tests online, implicit processing of information between vision, phonology and semantics in order to examine the effects of literacy on the language processing system, before presenting our model's design and results.

Changes to phonological representations and literacy

The aspect of speech processing for which there has been most exploration for an influence of literacy is in the domain of phonological awareness, defined as "one's degree of sensitivity to the sound structure of oral language" (Antony & Francis, 2005). There is substantial evidence indicating that, over the course of development, individuals become increasingly sensitive to smaller linguistic units within the speech signal. Children first gain awareness of larger units such as syllables before they are able to display an awareness of smaller units such as onsets and rhymes (Alcock, Ngorosho, Deus, & Jukes, 2010; Antony & Francis, 2005; Goswami, 2003). However, debate remains as to the cause of this improvement.

Firstly, what is the role of literacy acquisition? Is perceptual categorization of speech sounds dependent on reading acquisition (Burnham, 2003)? Does literacy lead to a finer tuning of perceptual categories and, consequently, improvements in the precision of phoneme identification (Hoonhorst, Medina, Colin, Markessis, Radeau, Deltenre & Serniclaes, 2011; Serniclaes, Ventura, Morais & Kolinsky, 2005)? Or does literacy not play a crucial role, instead is it that the fidelity of phonological representations increases across development driven by the need to differentiate, within an increasingly large lexicon, between an increasing number of phonologically similar items (Garlock, Walley & Metsala, 2001; Storkel, 2002)?

There is growing evidence that for (at least) explicit awareness of fine grain phonological units, individuals require exposure to alphabetic literacy training. Experiments that require children to make explicit judgements regarding a word's phonological structure show that children perform largely at chance prior to literacy training, however once engaged in training their performance on such tasks greatly improves (Alcock et al., 2010; De Jong & Van Der Leij, 2003; Hulme, Snowling, Caravolas & Carroll, 2005; Morrison, Smith & Dow-Ehrensberger, 1995; Treiman & Zukowski, 1991). Critically, similar tests have been conducted on illiterate adults with such individuals also displaying chance level performance on tasks requiring explicit phoneme manipulation or judgments (Adrian, Algeria & Morais, 1995; Loureiro, Braga, Souza, Queiroz, & Dellatolas, 2004; Morais, Cary, Alegria & Bertelson, 1979; Sciliar-Cabral, Morais, Nepomuceno, & Kolinsky, 1997). It has also been observed that late literates (individuals who learn to read in adulthood) although displaying improved performance compared to illiterates on phonological awareness tasks still perform worse than early literates (individuals who learn to read during childhood) (Morais et al., 1979; Morais, Bertelson, Cary & Alegria, 1986). Although performance of illiterates on phonological awareness tasks has been shown to be very poor, illiterates display improved performance (although performance is still lower than literates) on metaphonemic judgement tasks (Syllable Detection: Morais, Content, Cary, Mehler & Segui, 1989; Rhyme Awareness: Adrian, Algeria & Morais, 1995; Morais et al., 1986; Phonological Length: Kolinsky, Cary & Morais, 1987). Such data indicates that increases in phonological awareness, displayed by literate children, do not emerge simply as a result of greater exposure to spoken language or as the system matures, instead this evidence implicates literacy training as the critical factor in enabling explicit phonological awareness.

What is less clear is the impact of literacy on online speech processing. The above studies require participants to make explicit judgements regarding phonological properties of words. Based on this evidence alone it is not possible to say whether the progression towards explicit knowledge of more fine-grained components in the speech signal is also mirrored in an individual's implicit abilities when processing speech online.

Evidence for effects of literacy for online speech processing is less prevalent and less conclusive. Reis and Castro-Caldas (1997) observed that illiterates performed worse than literates on a pseudo word repetition task, whereas both populations performed equally well when repeating real words, suggesting that sub-lexical representations of spoken words were less readily accessible to the illiterate participants. Literacy has also been shown to influence categorical perception in speech. Serniclaes et al. (2005) showed that literates displayed sharper boundary precision in response to *ba-da* contrasts than illiterates, an effect that correlated with reading level (Hoonhorst et al., 2011). Such findings are consistent with an increase in the fidelity of phonological representation as a consequence of literacy, yet could instead indicate a more subtle refinement of categorical boundaries rather than confirming a prior absence of phoneme level representations (Burnham, 2003).

Although this evidence is largely consistent with literacy leading to more fine grained processing of the speech signal, it provides little insight regarding the stages in online speech processing affected by literacy training, for example does literacy lead to changes in early perceptual processing or are observed differences dependent on feedback from later activated orthographic knowledge? Such insight is important as phonological processing occurs rapidly with effects often transitory and dynamic in nature, so understanding the timing of these effects may provide the necessary evidence required to isolate differences in underlying cognitive processing.

Behavioural evidence is scarce regarding time-course effects of literacy, though one study that isolates timing differences (Ventura, Kolinsky, Querido, Fernandes & Morais, 2007) compared performance of literates and illiterates on a picture word interference task in which named pictures shared only the first phoneme with a spoken word. Results showed a phonological priming effect for both groups. However, illiterates only displayed an effect at later SOAs. This is compatible with more coarse grained processing of the speech input in illiterates, as it could be argued that more of the speech signal needs to unfold before overlapping representations are activated and can exert an influence on behaviour. ERP

(Event-related potential) data has also provided a productive means of probing time-course effects of literacy on online speech processing. Such studies demonstrate an early influence of orthography during spoken word processing, critically with effects observed in windows prior to points that are classically viewed as the time point of lexical access (Semantic categorisation task: Pattamadilok, Perre, Dufau, & Ziegler, 2009; Lexical decision task: Perre, Midgley, & Ziegler, 2009; Perre, Pattamadilok, Montant, & Ziegler, 2009; Perre & Ziegler, 2008).

Ziegler and Ferrand (1998) suggest that the mechanisms underlying the effects of orthography on online speech processing are that following literacy training orthographic representations are activated online when processing spoken words and it is such online activation that leads to effects of orthography on speech processing. Neuroimaging evidence consistent with this hypothesis can be found in Dehaene et al. (2010) in which online speech processing tasks in literates, but not illiterates, were observed to activate brain regions associated with orthographic processing. However, such evidence is not incompatible with an alternative restructuring hypothesis in which the process of learning orthographic mappings leads to adaptation in other language processing regions. For example, phonological processing regions may be restructured so that processing reflects characteristics of orthographic representations, such as being finer grained (Muneaux & Ziegler, 2004; Taft & Hambly, 1985; Taft, 2006). Neural data has also provided evidence in support of a restructuring account, by isolating effects of literacy on speech processing to regions associated with phonological processing. For example, Perre, Pattamadilok, Montant and Ziegler (2009) localized the source of orthographic consistency effects during spoken word recognition observed in ERP data to classic phonological processing regions (left BA40). Further, Pattamadilok, Knierim, Duncan and Devlin (2010) demonstrated that orthographic consistency effects during auditory lexical decision tasks can be removed when disturbing processing in these phonological processing regions (left supramarginal gyrus) using repetitive transcranial magnetic stimulation, while they were not affected by disturbance of orthographic processing regions (left ventral occipitotemporal cortex).

Psycholinguistic grain size theory (Ziegler & Goswami, 2005) offers a processing level model that connects exposure to the written forms of words to increased granularity of phonological processing. It is also largely consistent with the behavioural and neural data presented earlier. Grain size theory proposes that learning to map between orthographic and

phonological representations leads to a restructuring of phonological representations and is necessary to develop awareness of fine grained structure in the phonological lexicon, with the nature of the correspondence between orthographic units and phonological units within a given language determining the granularity of restructuring for that language. Ziegler, Bertrand, Tóth, Csépe, Reis, Faísca et al. (2010) found a relationship between phonological awareness and reading performance across a range of alphabetic orthographies in children in second grade of school. Though this relation was found to be stronger for more opaque orthographies, this may be because readers of transparent orthographies develop ceiling effects in phonological awareness skills earlier in reading exposure than readers of opaque orthographies (Caravolas, Volin, & Hulme, 2005). Nonetheless, training on orthographies, where the correspondence between individual phonemes and letters is largely consistent, as in the case of alphabetic languages, is likely to lead to finer-grained phonological representations, in comparison to orthographies where orthographic units correspond only to larger, coarser-grained phonological units comprising multiple phonemes, for example in logographic languages. Awareness of larger units within words (i.e. syllables, onsets, rhymes) may proceed without literacy training; however for awareness of fine grain units to emerge (i.e. phonemes) it has been proposed that explicit training is necessary. Evidence in support of this position comes from observed similarities in processing between illiterates and logographic literates, for example Chinese literates, where there is little systematic correspondence between orthographic representations and the sequence of speech sounds that constitute their spoken form (Brennan, Cao, Pedroarena-Leal, McNorgan & Booth, 2013; Cao, Khalid, Lee, Brennan, Yang, Li, Bolger & Booth 2011; Cheung, Chen, Lai, Wong & Hills, 2001; Ho & Bryant, 1997; Huang & Hanley, 1995, 1997; McBride-Chang, Bialystok, Chong & Li, 2004; Read, Yun-Fei, Hong-Yin & Bao-Qing, 1986; Shu, Peng & McBride-Chang, 2008).

In our model of online speech processing we test a phonological restructuring hypothesis consistent with psycholinguistic grain size theory, in which learning to map between orthographic and phonological representations leads to changes in the granularity of phonological processing that reflect the structure of the orthographic system on which the system is trained. Therefore, training on alphabetic languages, in which there is a regular mapping between individual orthographic and phonological units leads to more fine grained phonological processing.

Cognitive efficiency and literacy

The effects of literacy, however, have not been isolated only to the domain of phonological processing. Historically, illiteracy has been linked to reduced performance on a range of cognitive tasks, e.g., visual perception (Luria, 1976), reasoning (Levi-Bruhl, 1923), and memory (Vygotsky, 1978). However, isolating the role of literacy from other factors such as pre-existing cognitive deficits or increased exposure to formal schooling is a substantial challenge. Yet, more recent studies that have attempted to control for such factors have continued to demonstrate a link between literacy and changes in cognitive performance on tasks that extend beyond the domain of phonological processing.

Performance on standardized memory tasks has been observed to differ between literate and illiterate groups. Specifically, illiterates display worse performance than literates on digit span tasks, in which participants are required to repeat a sequence of digits (Reis, Guerriero & Petersson, 2003). In the domain of semantic processing, performance on semantic fluency tasks, in which participants are required to produce as many items from a pre-specified semantic category as possible, has also been shown to differ between literate and illiterate groups (Kosmides, Tsapkini, Folia, Vlahou & Kiosseoglou, 2004; Reis & Castro-Caldes, 1997), though see Da Silva, Petersson, Faisca, Ingvar & Reis (2004) for an alternative account of semantic effects being due to differences in general knowledge.

Effects on visual processing have also been observed and appear to extend as far as low level perceptual processing. For example in a recent study by Szwed, Ventura, Querido, Cohen and Dehaene (2012), illiterates performed worse than literates on a contour integration task, in which participants were required to indicate the direction of an image when the image was distorted by low level visual noise. In a visual target detection task in which participants were required to touch red squares placed among yellow squares on a computer screen, illiterates were shown to be slower and less accurate than literates (Bramao, Mendonca, Faisca, Ingvar, Petersson & Reis, 2007). However, such effects do not seem to be driven purely by low level perceptual differences. A more recent study examining visual search behaviour in literate and illiterate groups also observed slower performance in illiterate groups (Olivers, Huettig, Singh & Mishra, 2014), yet demonstrated that the observed difference in behaviour was largely accounted for by low literates needing more time between fixating the target and producing a required motor response. A possible explanation for this consistent reduction in performance displayed by illiterates across many cognitive

domains would be that literacy leads to a general increase in efficiency of cognitive processing.

General processing speed (Kail & Salthouse, 1994, Salthouse, 1996) has been shown to correlate with performance on a wide range of cognitive tasks (Kail & Salthouse, 1994; Li, Lindenberger, Hommel, Aschersleben, Prinz, & Baltes 2004; Salthouse, 2005), and has been proposed to be the mechanism of increased cognitive efficiency as a consequence of literacy training. For instance, Stoodley and Stein (2006) showed that literacy skills correlated with a general increase in speed of performance on a pure motor task. It has been suggested that general processing speed is related to the rate at which information propagates from one node in a network to another (Kail & Salthouse, 1994; Salthouse, 1988). Such arguments are consistent with recent research in the field of neuroscience, into the effects of myelination in the human brain. Measures of myelination and white matter integrity have been shown to be reflected in the efficiency (Deary, Bastin, Pattie, Clayden, Whalley, Starr & Wardlaw, 2006; Engel, Fries & Singer, 2001; Li, Liu, Qin, Li, Yu & Jiang, 2009) and the speed (Gutierrez, Boison, Heinemann & Stoffel, 1995; Madden, Bennett & Song, 2009; Penke, Maniega, Murray, Gow, Hernandez, Clayden, Starr, Wardlaw, Bastin & Deary, 2010; Tolhurst & Lewis, 1992; Waxman, 1980) of information processing, with such factors shown to modulate performance on a range of cognitive tasks (Deary et al, 2006; Li et al., 2009; Turken, Whitfield-Gabrieli, Bammer, Baldo, Dronkers & Gabrieli, 2008). Critically, myelination has been shown to be modifiable by experience (see Fields, 2008) and to increase as a result of learning (Bengtsson, Nagy, Skare, Forsman, Forssberg & Ullen, 2005) and therefore has the potential for modulation by environmental variables such as exposure to literacy training. Studies have indeed shown that myelination of brain regions associated with language processing coincides with vocabulary acquisition (Pujol, Soriano-Mas, Ortiz, Sebastian-Galles, Losilla & Deus, 2006).

Models of literacy effects on language processing

Many of the most influential cognitive models of speech processing do not implement a role for orthographic knowledge (e.g., Cohort Model, Marslen-Wilson & Tyler, 1980; MERGE, Norris, McQueen, & Cutler, 2000; Shortlist B, Norris & McQueen, 2008; TRACE, McClelland & Elman, 1986), however there is one model of which we are aware that provides insight into potential effects of literacy on phonological processing of words. Harm and Seidenberg (1999) compared behaviour of a computational model trained to generate

stable phonological representations of monosyllabic words, where the phonological representation was input as a set of phoneme features representing each phoneme in the word, to a model that in addition mapped orthographic representations onto the phonological representations, and assessed the effect of this literacy on the model's performance on a range of single-word phonological processing tasks. They found that when one phoneme within a word was affected by noise, the literate model was better able to restore the phoneme. They also found that the literate model represented words with the same rhymes as more similar in terms of the internal state of the model than the illiterate model.

These simulations provide an explicit description of how learning orthographic mappings can lead to emergent effects on phonological processing, modulating the componential nature of processing. They observed that changes to connection weights within the phonological network as a consequence of literacy training were greater for within-segment weights (within rime and within onset connections), rather than intersegmental weights (those crossing the onset-rime boundary). This they argue is consistent with evidence indicating that literacy leads to increased sensitivity to smaller phonological units. Harm and Seidenberg's (1999) model suggests that increased componentiality may be a consequence of literacy, consistent with the psycholinguistic grain-size theory and the restructuring hypothesis. However, the model represented phonology in terms of individual phonemes, thus increasing the chances that the model will discover phoneme-level representations. The model also did not test the potential effect of general cognitive processing advantages as a consequence of literacy within the model. For instance, similar observations of increased componentiality within the model could equally be a consequence of a model with greater fidelity of representations rather than (due only to) changes in the granularity of processing. For example, a model that possesses noisier representations is likely to perform worse on restoration tasks and may also represent words with the same rhymes as less similar.

Explicit and implicit phonological processing tasks

The above behavioural and computational studies provide substantial converging evidence for a connection between acquisition of literacy and the fidelity of phonological representations of words (Dijkstra, Roelofs & Fieuws, 1995; Chereau, Gaskell & Dumay, 2007; Hulme, Bowyer-Crane, Carroll, Duff, & Snowling, 2012; Kolinsky, Pattamadilok & Morais, 2012; Ventura, Morais, Pattmadilok & Kolinsky, 2004; Ziegler and Farrand, 1998). However, these previous studies have generally focused on explicit tasks addressing

manipulations of the phonological forms of isolated words, and there is little extant evidence for behavioural consequences that may result from differences in phonological activation during online speech processing. This is important ecologically because these previous studies have focused either on manipulations of the phonological representation itself, or on the extent to which phonological representations are similar to one another, rather than the use that the language processing system makes of these representations. Language processing involves combinations of phonological, orthographic, and semantic representations in interaction with sensory input about the environment, and so determining the effects of literacy on language processing should take this complexity into account, rather than only focusing on one small aspect of the language processing system. Implementing the complexity of the system also permits testing the various accounts of effects of literacy on representations other than only phonological forms, which may prove important for distinguishing competing accounts based on cognitive efficiency or grain-size of phonological representations.

One way in which use of phonological representations can be studied is through the visual world paradigm (Cooper, 1974, Tanenhaus et al., 1995). In studies of this kind, participants are presented with a visual display while simultaneously hearing a spoken utterance, and as these events unfold their eye gaze is recorded. The paradigm has been previously used to examine integration of information between visual and linguistic representations by manipulating the relationships between items in the visual scene and words presented in the auditory stimulus (for review, see Huettig, Rommers, and Meyer, 2011).

Huettig, Singh, and Mishra (2011) conducted a visual world paradigm study with two populations in India both of which were native Hindi speakers (all materials were in Hindi). One was a high literate population, comprising undergraduate university students, and the other was a low literate population, that complied with the UNESCO definition of illiterate (provided earlier in this paper) yet who were fully integrated within Indian society. Low literates were employed and displayed no obvious social, cognitive or neurological deficits. The critical difference between populations was the amount of exposure to formal education. In their experiment 1, participants listened to sentences containing a target word and were shown scenes containing a semantic competitor (item that shares a subset of its semantic features with the spoken target word; e.g. target = beaker, semantic competitor = fork), a phonological onset competitor (item that shares its phonological onset with the spoken target word; e.g. target = beaker, phonological competitor = beetle) and two unrelated distractors

(items that did not share either a phonological or semantic relationship with the spoken target word). In their experiment 2, scenes only displayed a phonological competitor and three unrelated distractors. On experimental trials visual scenes did not contain the named item, whereas filler trials contained scenes displaying pictures of the spoken target word in addition to three unrelated distractors. In both experiments participants performed a look and listen task, this simply required participants to look at the scenes while listening to the spoken sentence with no additional explicit task.

The results of Huettig et al.'s (2011) two experiments are displayed in Figure 1. When presented with scenes containing phonological onset competitors and semantic competitors, high literates looked first at phonological competitors and then later at semantic competitors once information within the unfolding speech mismatched with the name of the phonological competitor, replicating earlier research (Huettig & McQueen, 2007). Low literates on the other hand only displayed increased fixation of semantic competitors, at no point fixating phonological competitors consistently more than unrelated distractors. Also, overall fixation of semantic competitors by low literates was lower than that displayed by high literates. In the second experiment, when participants viewed scenes containing only a phonological onset competitor accompanied by unrelated distractors, high literates again displayed a pattern of fixation toward the phonological competitor that tightly mirrored the phonological overlap in the speech signal. Low literates on the other hand, unlike in experiment 1, did display increased fixation of the phonological competitor compared to unrelated distractors, but in contrast to high literates, their fixations of phonological competitors were not tightly time locked to the unfolding speech signal. Low literates looked marginally more at the phonological competitor over the first 1000ms post word onset but did not display the rapid increase and decrease in looks towards this category of item in response to signal overlap as shown by high literates.

These results support a qualitative difference between high and low literate populations in their use of phonological information, whereas only a small quantitative difference in terms of semantic processing. In terms of the two theories of effects of literacy on language processing – the fine-grained phonological representation and the cognitive efficiency theories – the results may be consistent with either. For high literate participants, they are sensitive to the phonological overlap between representations of words at the point at which the phonology of competing visual objects applies. For low literate participants, such

sensitivity is not observed suggesting that fine-grained phonological distinctions between words are a consequence of literacy. The low literates may be instead biased toward mapping at a semantic level and only map at a phonological level when pushed by restrictions on the nature of the representational overlap in the environment. Huettig et al. (2011) conjecture that literacy results in these effects by strengthening existing phonological representations and providing a tighter coupling between phonological representations activated by the visual environment and phonological representations activated by spoken language input.

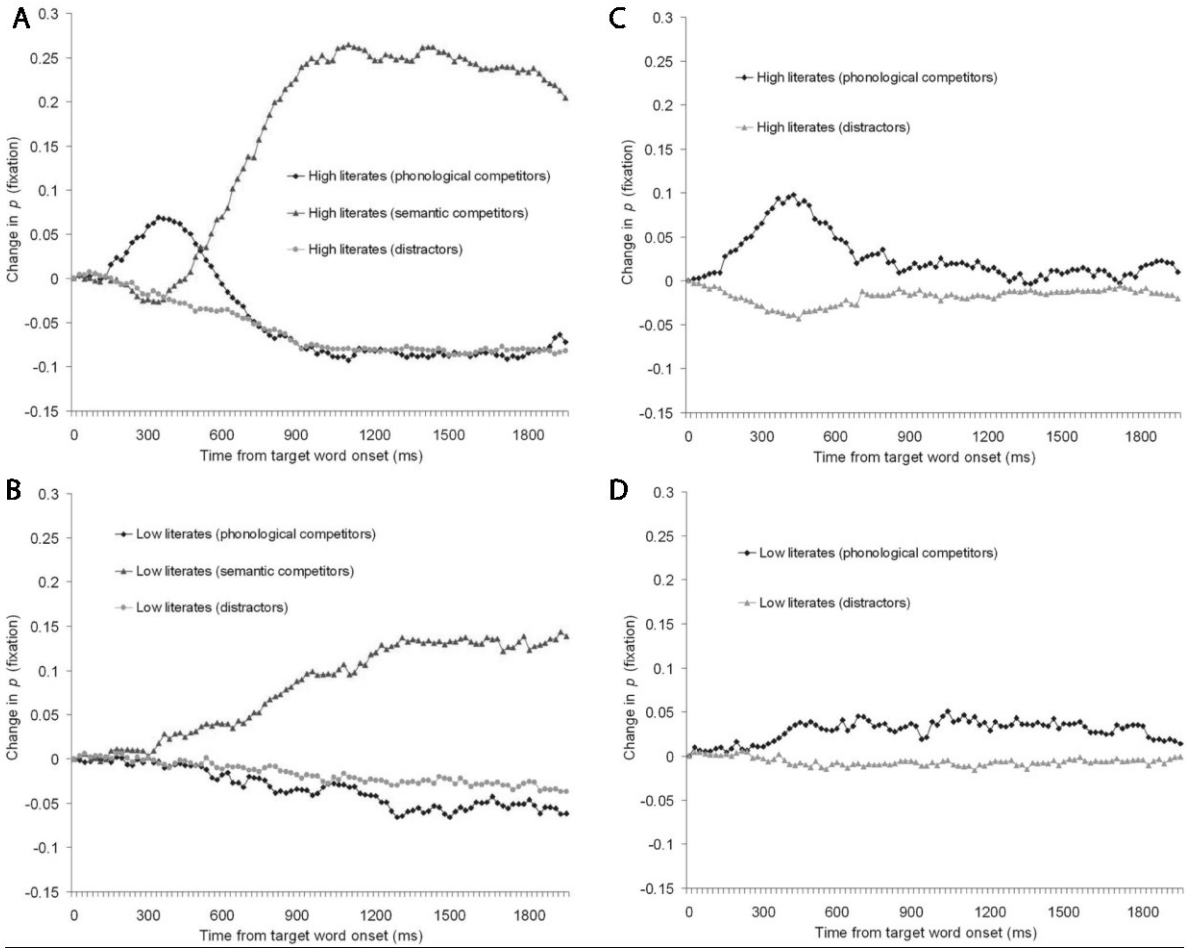

Figure 1: Results of Huettig, Singh and Mishra (2011). Charts display the change in fixation proportions for high (A and C) and low (B and D) literates when presented with scenes containing either a phonological competitor, semantic competitor and unrelated distractors (A and B), or a phonological competitor and unrelated distractors (C and D). (Figures as published in Huettig, F., Singh, N., & Mishra, R. K. (2011). Language-mediated visual orienting behaviour in low and high literates. Frontiers in Psychology, 2, 285. doi:10.3389/fpsyg.2011.00285.).

However, an alternative perspective is that the results are a consequence of greater efficiency in cognitive processing. Turken et al. (2008) highlight the efficiency of signal transmission across white matter tracts as a particularly significant factor in determining performance on tasks that require the complex integration of information from multiple operations. Therefore, the processes driving language mediated eye gaze, in which information from auditory, visual, semantic and eye gaze processing regions must be tightly integrated and used to coordinate behaviour, may be greatly influenced by such a variable. In terms of this theory, the observed results are then just due to the greater effectiveness and fidelity of the phonological representations for the high literate participants, rather than a qualitative difference in the grain-size of the processing.

The results of Huettig et al. (2011) alone do not provide a means of testing these two alternative hypotheses for the effects of literacy training on online language processing. However, the current study aims to demonstrate that through combining the rich behavioural data provided by the Visual World Paradigm with a computational model of language meditated visual attention it is possible to gain traction on such questions of effects of literacy on language processing that to date have eluded researchers.

The visual world paradigm is well suited to the investigation of phonological processing effects as it enables researchers to test effects not only of phonological processing but also of semantic effects on processing. Capturing these two effects in a single online task is required in order to distinguish between these alternative theories. Further, the use of a dense sampling method such as eye gaze, to compare between literate and illiterate behaviour, allows examination of moment by moment performance providing additional constraints on underlying processing differences during online speech processing that would otherwise be lost in more global measures. Within the current modelling approach, we use this rich behavioural measure to compare alternative theoretical explanations for the observed data by testing the behaviour of explicit implementations of each theory in a computational model.

In order to test these theories of literacy effects on language processing, we adapted the multimodal integration model (MIM) of language mediated visual attention (Smith, Monaghan & Huettig, 2013a; 2013b). This model offers an explicit description of the information and processes that drive complex multimodal behaviour in language processing. The model has previously been shown to replicate a broad range of word level effects, displayed by literate populations, reported in the visual world literature (see table 1). For

example, the model successfully replicates contrasts across modalities in the effect of representational overlap on fixation behaviour of literate populations. The model replicates semantic effects observed in the visual world paradigm (Huettig, Quinlan, McDonald & Altmann, 2006; Mirman & Magnuson, 2009; cf. Yee & Sedivy, 2006) in that it fixates items that share a semantic relationship at levels proportional to the number of semantic features shared between items. In contrast fixation of items that share purely phonological relationships is, in addition, dependent on the temporal location of overlapping phonological features. The model also replicates the difference between effects of phonological rhyme and phonological onset overlap as reported in Allopenna, Magnuson and Tanenhaus (1998), fixating items that share initial phonemes earlier and with increased probability than items that share phonemes in final positions.

This model of language mediated visual attention has been shown to be not only sufficient to account for the experimental effects of the visual world paradigm for literate participants but also, importantly for this investigation, demonstrates how such features of behaviour are emergent properties of both the structure of representations and the computational properties of the mappings performed between them. The model achieves this through capturing both the process through which this behaviour is acquired and through use of a parsimonious architecture that implements only minimal assumptions about processing mechanisms. The model therefore is appropriate for testing the impact of differences between populations in representational structure on eye gaze in the visual world paradigm.

Previous computational models of the effects of phonological processing used in conjunction with visual world data have tended to model processing in a single modality (Allopenna et al, 1998; Mirman, Dixon & Magnuson, 2008), the model used in this study however captures the processing of phonological, semantic and visual information. The model provides as an output a dynamic measure of the location of fixation across multi-object scenes over time, which is dependent on the integrated processing of information across all three modalities. The model also differs from previous models of effects on phonological processing where phonology has been used as both an input and output measure (Harm & Seidenberg, 1999). In contrast the chosen model has a different dependent variable ‘eye gaze’ therefore phonological manipulations are only indirectly related to behaviour. The ability to detect simultaneously the effect of phonological and semantic influences on performance permit greater discrimination of the effects of the phonological representation or the cognitive

efficiency theory of literacy by investigating whether each implemented theory matches the effects of literacy in both representational domains.

Table 1: Table presenting Visual World data successfully replicated by the MIM model of language mediated visual attention (Smith, Monaghan & Huettig, 2013a; 2013b). The items displayed within scenes in each empirical study are listed with observed competitor effects highlighted in bold. Visual = visual competitor, Semantic = semantic competitor, Onset = phonological onset competitor, Rhyme = phonological rhyme competitor. [a] *= Study presented near and far semantic competitors on separate trials.* [b] *= Experiment 1.*

Study		**Scene**			
Authors	**Year**	**Item 1**	**Item 2**	**Item 3**	**Item 4**
Allopenna et al.	1998	Target	**Onset**	**Rhyme**	Distractor
Dahan & Tanenhaus	2005	Target	**Visual**	Distractor	Distractor
Huettig & Altmann	2007	**Visual**	Distractor	Distractor	Distractor
Yee & Sedivy	2006	Target	**Semantic**	Distractor	Distractor
Huettig & Altmann	2005	**Semantic**	Distractor	Distractor	Distractor
Mirman & Magnuson[a]	2009	Target	**Near Sem**	**Far Sem**	Distractor
Huettig & McQueen[b]	2007	**Onset**	**Semantic**	**Visual**	Distractor

We manipulate both grain-size in phonological processing as well as processing efficiency within this neural network model of language mediated eye gaze (Smith et al, 2013a). As finer grained representations more clearly encode the regions of representations that differ or overlap, we predict that increasing the granularity of phonological processing will increase the salience of phonological onset competitors at points in which the phonology of the unfolding spoken word overlaps with the phonological representation corresponding to the phonological onset competitor and reduce its salience at points in which the signal mismatches. Therefore a system that processes phonological information at a finer grain size will be able to use such information to distribute attention more dynamically in response to information in the unfolding auditory input.

We also predict that semantic effects on the other hand should not be affected significantly by the granularity of phonological processing. In previous computational simulations of language mediated eye gaze (Smith et al., 2013a), semantic effects are driven by overlap between the semantic representations of the visually depicted object and the spoken word. Overlap effects are therefore dependent on the level of activation of overlapping semantic features triggered by the concurrent visual and auditory input. As more of the phonological signal unfolds, the activation of corresponding semantic features will increase. When the words' semantic properties are maximally activated, if the level of representational overlap does not differ across distinct phonological grain sizes, then the literate and illiterate simulations should activate semantic representations equally.

In addition, we predict that manipulations of processing efficiency will have a greater impact on semantic effects than phonological effects. Within the model, activation of semantic information is more sensitive to the efficiency of information transfer within the network as it is not directly activated by the visual or auditory input, but instead activated as a consequence of signals that flow through the network from phonological and visual input layers. We also predict that such effects will be quantitative rather than qualitative in nature as the structure of signal overlap will not differ but simply lead to an overall reduction in the activation of overlapping features, which in turn will result in a quantitative reduction in the saliency of semantic competitors.

2. Method

Architecture

The neural network model used within this paper is based on the MIM model of language mediated eye gaze presented in Smith et al, (2013a). The same network architecture (see Figure 2) was used for all simulations. The model consisted of four modality specific processing layers connected via a central resource. We know from behavioural data recorded in visual world studies that language mediated visual attention is driven by the interaction of information extracted from the visual environment and speech signal in terms of semantic, visual and phonological representations (e.g., Huettig & McQueen, 2007). The MIM architecture offers a parsimonious solution to how these modalities interact, and a means by which the emergent properties of this complex interaction can be harnessed. The architecture allows competition at multiple levels of representation, parallel activation of representations,

the integration of information from multiple modalities and allows for both inhibitory and excitatory associations between representations, all of which have been proposed as important theoretically for reflecting behavioural effects in the Visual World Paradigm (see Smith et al., 2013a, for review). The philosophy of the model was to investigate the extent to which information from different modalities interacts in language processing tasks, with constraints within the model's processing resulting from the nature of the mappings between representations, rather than imposed architectural assumptions.

The visual layer (80 units) simulates the extraction of visual information from up to four locations in the visual field. The layer is divided into four 20 unit slots. Each slot encodes the information available at a single location in the visual field, with active units representing visual feature information about objects. The phonological layer (60 units) simulates the temporally unfolding speech signal. It comprises six time slots, each of which contains 10 units which encode the phonological features of the auditory input at a given time point. Phonological layer time slots are activated sequentially during the presentation of the spoken word to the model. For example, at word onset the activation pattern corresponding to the initial information of the unfolding spoken word will be presented to the first 10 units in the phonological layer, while all other units in the layer remain inactive. At the next time step the activation pattern corresponding to the second portion of the spoken word will in addition be presented to units 11-20 in the phonological layer, while again all subsequent units will remain inactive. This process continues, with an additional phonological portion presented at each subsequent time step until the entire phonological representation of the spoken word is presented to the phonological layer (i.e. all six portions corresponding to the six time slots). The semantic layer (160 units) represents the semantic features for items presented either in spoken or visual object form. Finally, the eye layer (4 units) provides a measure of the model's direction of gaze across the four possible locations in the visual field, with each unit in the eye layer associated with one of the four quadrant locations in the visual field. Activation of eye layer units was taken to correspond to the probability of fixating associated locations in the visual environment, with the location associated with the most highly activated eye layer unit interpreted as the location currently fixated. All four layers were fully connected to a central integrative layer (400 units) which was fully self-connected and also fully connected to allow activation to feed back to eye and semantic layers.

At each time step activation passed between all layers in the network (see appendix). Training trials extend over a total of 14 time steps to enable activation to cycle between representations in the model. During testing, this period was extended to allow for insight into the time-course of interaction between representations across modalities within the model.

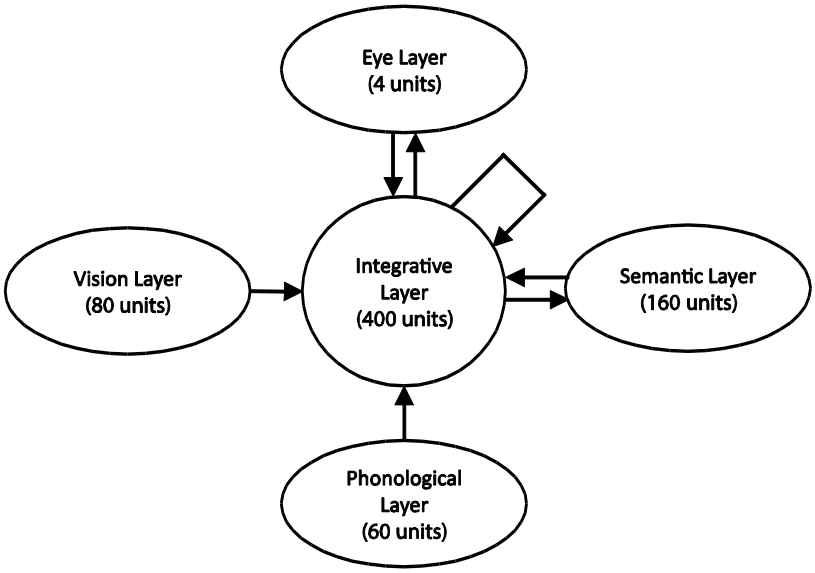

Figure 2: The MIM model of language mediated visual attention

Representations

To ensure the overlap for visual, phonological, and semantic representations was fully controlled both within and across modalities, a fundamentalist approach was taken in their construction, i.e., that "a model should embody only the principles that are theorized to account for the phenomenon in focus" (Kello & Plaut, 2000). Consequently, artificial corpora were constructed each consisting of 200 items (or words in the artificial corpus), with each item defined by a unique visual, a unique semantic and a unique phonological representation. We constructed 15 versions of the corpus for each manipulation of grain size, though corpora were matched across cognitive efficiency manipulations, resulting in 45 artificial corpora (3 grain sizes x 15 instantiations of each grain size). This was to ensure that chance variation in distribution of the spoken forms of words, and chance random starting states of the model, did not bias model performance.

Visual representations were 20 unit binary vectors with each unit representing the presence or absence of a distinct visual feature. Visual representations were composed of two distinct

components, one component representing coarse (low frequency) visual features and the other fine (high frequency) visual features. Visual features were randomly assigned to items with p(active) = 0.5. Ten units per component meant that encoding of 200 unique representations was possible without populating the representational space to maximal density, ensuring that visual representations could be distinguished by one or more visual features.

Semantic representations were designed to reflect discrete and relatively sparse semantic feature-based representations for words (Harm & Seidenberg, 2004). For each item a unique set of 8 from a possible 200 semantic features were activated, with features randomly assigned.

Table 2: Cosine distance between target and distractor representations. Records mean [$\bar{x}$] and standard deviation [σ] calculated across all items and corpora (15 per grain size).

Modality	**Simulation**	**Distractor**	**Cosine Distance [$\bar{x}$,(σ)]**		
Phonological			**Onset**	**Rhyme**	**Overall**
	Fine Grain	Competitor	0.00 (0.00)	0.51 (0.12)	0.34 (0.08)
		Unrelated	0.51 (0.16)	0.49 (0.11)	0.50 (0.09)
	Moderate Grain	Competitor	0.17 (0.10)	0.42 (0.09)	0.33 (0.07)
		Unrelated	0.51 (0.14)	0.51 (0.10)	0.51 (0.08)
	Coarse Grain	Competitor	0.35 (0.12)	0.34 (0.09)	0.34 (0.07)
		Unrelated	0.51 (0.14)	0.50 (0.10)	0.50 (0.08)
Semantic			**Overall**		
	All Simulations	Competitor	0.50 (0.00)		
		Unrelated	0.96 (0.06)		

Representational overlap between items was controlled in visual, semantic and phonological dimensions (see table 2). Embedded within each corpus were 40 target items. Within the corpus were also embedded for each target item a semantic competitor (n = 40) and a phonological competitor (n = 40). Competitors shared increased representational overlap with their assigned target item in a single modality (either semantic or phonological). Semantic competitors shared 4 of 8 semantic properties with their assigned target, while all unrelated items shared a maximum of 1 semantic property with any other item.

Simulating differences in phonological processing

To simulate differences in the granularity of speech processing three forms of processing of the phonological input were constructed: fine, medium, and coarse grained phonological representations, reflecting variations in the componentiality of the phonology, from single phonemes up to a holistic word-level representation. Fine grained representations simulated phonological processing in which the unfolding speech signal activates a componential sequence of phonemes, and this was the representation used in previous simulations of standard speech processing with the MIM model (Smith, Monaghan, & Huettig, 2013). For the fine grained simulations, an inventory of 20 possible phonemes was constructed. Each phoneme was encoded as a unique 10 unit binary vector with units assigned with p(active) = 0.5, and each unit representing a phonological feature. For each word, a unique sequence of 6 phonemes was constructed by randomly sampling from the set of 20 phonemes. For the overlapping target-competitor pairs, the target and competitor shared their initial 2 phonemes (in Huettig, Singh & Mishra, 2011 phonological competitors shared at least the first two phonemes with the target word). No other word pairs within the corpus shared their initial two phonemes and no two words shared their initial three phonemes. These constraints applied equally to all the grain size conditions in order to ensure that comparisons between the grain sizes were controlled. However, the overall amount of overlap between items in the whole corpus is less than that observed in natural language, where overlap even up to the first 5 phonemes of 6 phoneme words is observed. This minimal overlap may have enhanced the size, rather than the qualitative nature, of phonological overlap observed in the results. All items had a unique sequence of the final three phonemes and no item contained more than two of the same phonemes in its entire representation.

For moderate grained representations two components encoded each word, analogous to syllable level representation of the stimulus. Each component was a unique 30 unit binary vector with units assigned with p(active) = 0.5. For the phonological overlap items, the target and competitor shared 2/3 of the features present in the initial component, and the remaining features in both the initial and second component were either shared or distinct with p = 0.5. The number of overlapping features was thus controlled across the fine and medium grained representations for phonologically overlapping items (see table 2), though they differed in terms of the componentiality of the overlap. Unrelated items shared all features with p = 0.5. Phonological similarity was thus simulated as distributed over the first half of the item, rather

than, as in the case of the fine grained simulations, with precise similarity over the first two phonemes of the item.

For coarse grained representations each word was defined by a single component, simulating a word-level phonological representation. This component was a unique 60 unit binary feature vector with p(active) = 0.5. For the overlapping pairs of target and competitor items, 1/3 of the features were identical between the two items, and the remaining features were shared with p = 0.5. Thus, the total number of similar features was the same as for the overlapping items in the fine and medium grained representations. Unrelated items shared features with p = 0.5.

To simulate the time-course of the unfolding phonological input, for all simulations sequences of 10 features of the spoken input were gradually presented. Thus from word onset an additional 10 features of the spoken input were presented at each subsequent time step, until by word onset + 5 time steps, all 60 features that define an item's spoken representation were presented as input to the phonological layer. This controlled for the temporal presentation of the auditory signal for each grain-size encoding, but the grain size of representations differed in terms of the componentiality of the presented features. Thus, for the fine grained representations, the presentation was phoneme-by-phoneme, whereas for the moderate grained representation, the presentation partially unfolded the syllable across three time steps, such that after three time steps the model was exposed to a full component. However, early stages of presentation did provide the model with information about the general sound of the syllable (so two similar syllables would have similar representations in the first ten features). For the coarse grained representation, again the presentation partially unfolded the word across six time steps, with similar sounding words having similar representations as the spoken form unfolded. Although, during language processing listeners would receive the same auditory signal with identical temporal properties, the grain size at which an individual processes speech will determine their ability to identify the components of words. Hence, when an onset competitor shares its initial two phonemes with a spoken target word (but not the entire syllable) a system processing at a finer grain will be quicker to detect the speech sound overlap than a system processing at a coarser grain. The fine, moderate and coarse grain representations implemented within this model capture this assumption.

Simulating differences in cognitive efficiency

There are several ways to simulate variation in cognitive efficiency in neural network models, including addition of noise to activations of units, reductions in processing resources to form mappings between representations, or reduction in overall levels of activation within the network (Harm & Seidenberg, 1999; Monaghan & Shillcock, 2004). We chose to simulate cognitive efficiency in the current models in terms of addition of noise. This decision was made so as to link the implementation of cognitive efficiency as closely as possible to theories about the information processing effects of literacy on neural processing, in terms of myelination of highly-trained pathways reducing noise levels in neural transmission via faster processing of high quality information. Gaussian noise was thus applied [$N(\mu=0, \sigma^2=0.02)$] to the output of all units in the network for the lower cognitive efficiency simulation, but no noise was added to the higher cognitive efficiency version[1]. This resulted in differences in the fidelity with which information passed through the network, and consequently the speed at which activation could accumulate in different modality layers in the model. Pilot simulations were used to establish an appropriate level of noise for the lower cognitive efficiency simulation. Simulations trained with noise sampled from [$N(\mu=0, \sigma^2=0.05)$] failed to learn some of the mappings between modalities that were a precursor to testing experimental performance of the model against the behavioural data and simulations trained with noise sampled from [$N(\mu=0, \sigma^2=0.01)$] displayed negligible differences in performance from simulations in which no noise was applied.

For each simulation set (fine grain, low efficiency; fine grain, high efficiency; moderate grain, low efficiency; moderate grain, high efficiency; coarse grain, low efficiency; coarse grain, high efficiency), 15 versions were trained on one of the distinct corpora and each was initialised with a different random seed.

Training

For each grain size (fine, medium, and coarse) we manipulated the model's cognitive efficiency (high, low) leading to a total of six sets of parameters for the model, and as mentioned earlier, there were 15 different simulation runs for each of these parameterisations.

[1] *The cognitive efficiency hypothesis as implemented within this study does not argue for a reduction of cognitive efficiency within the entire cognitive system. Instead, based on the argument of increased myelination in heavily trained networks, differences in cognitive efficiency would only result in networks that experience a difference in levels of training as a consequence of literacy acquisition. Within the model we only model such networks and therefore manipulation of global cognitive efficiency within the model is a valid reflection of the neural network changes associated with literacy.*

All simulations were trained on four tasks (see table 3) that aimed to simulate the tasks performed by participants in the natural learning environment. We assume that participants gain knowledge of an item's visual, semantic and spoken form by repeated and simultaneous exposure to these multiple forms of representation: It is through such experience that individuals acquire the associations between representations across modalities that later drive the behaviour observed in the laboratory setting.

Table 3: Procedure for the model's training trials.

Task	Vision		Phonological		Semantic		Eye	
	Description	Time Step	Description	Time Step	Description	Time Step	Description	Time Step
1. Vision to Semantics	4 visual representations randomly selected from the training corpus, 1 of which is randomly selected as a target.	0-14	Random time variant noise provided as input.	0-14	Target's semantic representation provided.	3-14	Target location fully activated, all other locations inactive.	0-14
2. Phonological to Semantics	Random time invariant noise presented to all visual input slots.	0-14	Target speech signal provided as staggered input	0-14	Target's semantic representation provided.	5-14	No constraints on activation	
3. Phonological to Location	Identical to procedure in task 1 and 4.	0-14	Identical to procedure in task 2.	0-14	No Constraints on activation		Target location fully activated, all other locations inactive.	5-14
4. Semantics to Location	Identical to procedure in task 1 and 3.	0-14	Identical procedure in task 1.	0-14	Semantic representation of target provided.	0-14	Target location fully activated, all other locations inactive.	2-14

Vision to semantics

This task aimed to simulate the learning that occurs during events in which individuals simultaneously view an item and determine some of its semantic properties, e.g., its function: seeing a fork and determining its use for eating. This was simulated by first randomly selecting four items from the artificial corpus, one of which was randomly selected as a target. The visual representations of each of the four items was then presented to the model at trial onset (time step 0), with each item randomly assigned to one of the four locations in the model's visual field. The eye unit relating to the location of the target's visual representation was also fully activated at trial onset with all other eye units fixed at zero activation. These values remained fixed for the remainder of the training trial. Throughout the trial small levels of variable background noise were provided as input to the phonological layer, simulating ambient background sound. Once sufficient time had passed allowing for activation to flow from the visual and eye layers to the semantic layer (i.e. time step 3) the item's semantic

representation was provided as a target and error was backpropagated through the network up to time step 14.

Phonology to semantics

This aimed to simulate the learning that takes place when an individual is exposed to an item's spoken form and is required to determine its semantic properties, for example, when hearing the word fork and eating from a fork. To simulate such occurrences, an item was first randomly selected from the training corpus and assigned the role of target. At trial onset the first 10 features of its phonological representation were presented in the initial slot of the phonological layer. At each subsequent time point a further 10 features of the target's representation were presented in the corresponding phonological input layer slots until the entire representation had unfolded. This remained present until the end of the training trial. Throughout such trials, random background noise was presented to the visual layer to simulate ambient stimulation of the visual system. Once the entire word had unfolded and sufficient time had elapsed for a signal discriminating the target from possible competitors to pass to the semantic layer (time step 5), the item's semantic representation was presented as a target and error was backpropagated until time step 14.

Phonology to vision

Orientating to an item when hearing its spoken form was trained by first selecting four items randomly from the training corpus and selecting one as a target. The visual representation of all four items was presented to the visual input layer at trial onset with locations in the visual field randomly assigned. Also coinciding with trial onset, the phonological representation of the target item began to unfold, with an additional 10 features of the target's phonological representation presented at each subsequent time step. Once the entire word had unfolded and sufficient time had passed to allow a discriminating signal to reach the eye layer (time step 5), the training signal was provided and error backpropagated. The training signal consisted of fully activating the eye unit relating to the location of the target's visual representation while fixing activity in all other eye units to zero.

Semantics to vision

A similar procedure was applied when training the model to orientate to the location of a target when provided with its semantic representation. Again four items were randomly selected from the training corpus and one randomly assigned as the target. The visual representations of all four items were presented to the visual input layer at trial onset, with

locations randomly assigned. Also coinciding with trial onset, the semantic representation of the target was presented. Throughout such training trials small levels of variable noise were provided as input to the phonological layer to simulate auditory background noise. Once sufficient time had elapsed for the signal from both visual and semantic layers to pass to eye layer units (time step 2), the training target signal was provided and error backpropagated. The training signal consisted of fully activating the eye layer unit associated with the location of the target's visual representation with zero activation in all other eye layer units.

All training tasks were randomly interleaved. In the natural language learning environment, items around the child are frequently left unnamed (Yu & Ballard, 2007). Hence, we assume during training that items based on their semantic properties are selected more frequently than items based on their spoken form, and so phonology to vision tasks were four times less likely to occur in training than other training tasks.

Connection weights were initialised with random weights taken from a uniform distribution [-0.1, 0.1]. Weights were adjusted online during the training process using recurrent back-propagation with learning rate 0.05 (see appendix). All simulations were trained on 850,000 training trials as this provided sufficient exposure for simulations to perform accurately on all four training tasks.

3. Results

Pre-test

Post-training, all simulations were tested on their ability to perform each of the four training tasks. Table 4 presents the accuracy of simulations on each task, averaged across 15 instantiations of each simulation. For tasks requiring the model to reproduce the semantic representation of the target when presented with its visual representation, simulations were tested on their ability to perform this task with the target presented in all possible locations in the visual field. Similarly, on orientation tasks, simulations were tested on their ability to orientate to the location of the target when the target was positioned in each of the four possible locations in the visual field. For phonological to semantic mapping tasks the model was tested four times on each item. Two measures were used to assess performance on each training task. For semantic mapping tasks we calculated the cosine distance between mean activation in the semantic layer during test trials and the semantic representation of all items

within the training corpus. In table 4, mean accuracy indicates the proportion of test trials for which the target's semantic representation was closest in terms of cosine distance to activation within the semantic layer. The second measure indicates the proportion of items within the training corpus for which activation in the semantic layer was closest to the semantic representation of the target in at least three out of four test trials. Two measures were also collected to assess model performance on orientation tasks. Mean accuracy indicates the proportion of test trials in which the eye unit relating to the location of the target's visual representation was most highly activated. The second measure provides the proportion of items within the training corpus for which the eye unit relating to the location of the target's visual representation was most highly activated on 3 out of 4 test trials.

Table 4: Trained model's performance on training tasks.

Simulation		**Task**							
Efficiency	**Grain**	**Visual to Semantic [b]**		**Phonological to Semantic [b]**		**Phonological to Location [a]**		**Semantic to Location [a]**	
		mean	**%**	**mean**	**%**	**mean**	**%**	**mean**	**%**
High	**Fine**	0.88	0.98	1.00	1.00	0.91	0.96	0.92	0.98
Low	**Fine**	0.85	0.97	1.00	1.00	0.90	0.97	0.92	0.99
High	**Moderate**	0.86	0.97	1.00	1.00	0.91	0.97	0.92	0.98
Low	**Moderate**	0.83	0.97	1.00	1.00	0.90	0.96	0.93	0.99
High	**Coarse**	0.87	0.98	1.00	1.00	0.91	0.97	0.93	0.99
Low	**Coarse**	0.82	0.97	1.00	1.00	0.90	0.96	0.93	0.99

([a] Location task: mean = mean proportion of test trials in which eye unit corresponding to target location was most highly activated, % = proportion of items within corpus for which target location was most highly activated on at least 3 of 4 test trials; [b] Semantic task: mean =proportion of test trials for which semantic layer activation was closest to target's semantic representation, % items = proportion of items for which semantic layer activation is closest to target's semantic representation on at least 3 of 4 test trials.)

The measures of model performance presented in table 4 demonstrate that all simulations were able to complete each of the four training tasks with a high degree of accuracy and displayed comparable levels of performance. There are significant differences between high and low cognitive efficiency simulations in task performance for visual to semantic mappings for the mean activation but not the % items measure, indicating that cognitive efficiency

drives a small difference in mapping location to semantic representations, as was predicted, but does not affect mapping from semantics, nor mappings from other modalities to semantics.

Simulating the effects of grain size and cognitive efficiency on language mediated visual attention

To simulate the conditions under which participants were tested in Huettig, Singh and Mishra (2011), Experiment 1, we presented the model with scenes containing a semantic competitor, a phonological onset competitor and two unrelated distractors while simultaneously presenting the phonological representation of the given target word. This was achieved using the following procedure. At trial onset the visual representations of the four items within the scene were presented. After a short delay, to enable pre-processing of the visual information, the spoken target word began to unfold (time step = 5), with an additional component of the target word's phonological representation presented at each subsequent time step until the full representation was revealed. The model's "gaze", was interpreted as being directed towards the location in the visual display associated with the most highly activated unit in the eye layer. 'Gaze' was recorded in this manner at each time step throughout the test trial, which lasted in total 29 time steps. For each instantiation of each simulation there were in total 960 test trials, with each item (n=40) occurring with competitors in all possible spatial configurations (n=24). Figure 3 displays the change in p(fix) from word onset (time step = 5), for each simulation (Figure 3A: High cognitive efficiency simulations; Figure 3B: Low cognitive efficiency simulations), with p(fix) recording the Luce ratio of fixations for each category of item within displays, averaged over all test trials (n=960) and instantiations (n=15).

We used linear mixed effects models (Baayen, Davidson, & Bates, 2008; Barr, 2008; Jaeger, 2008) to analyse the influence of grain size and cognitive efficiency on differences between competitor and distractor fixation. As a baseline for behaviour, we identified a preview time window (time steps 0 – 7), the period from trial onset until the first time point in which a signal discriminating between target and competitor within the phonological input can influence eye layer units. We then compared this baseline eye gaze performance of the model to its behaviour in time windows after this point, where the information from the phonological input met the visual system, distinguishing between early (time steps 8-18), and late (time steps 19-29) processing.

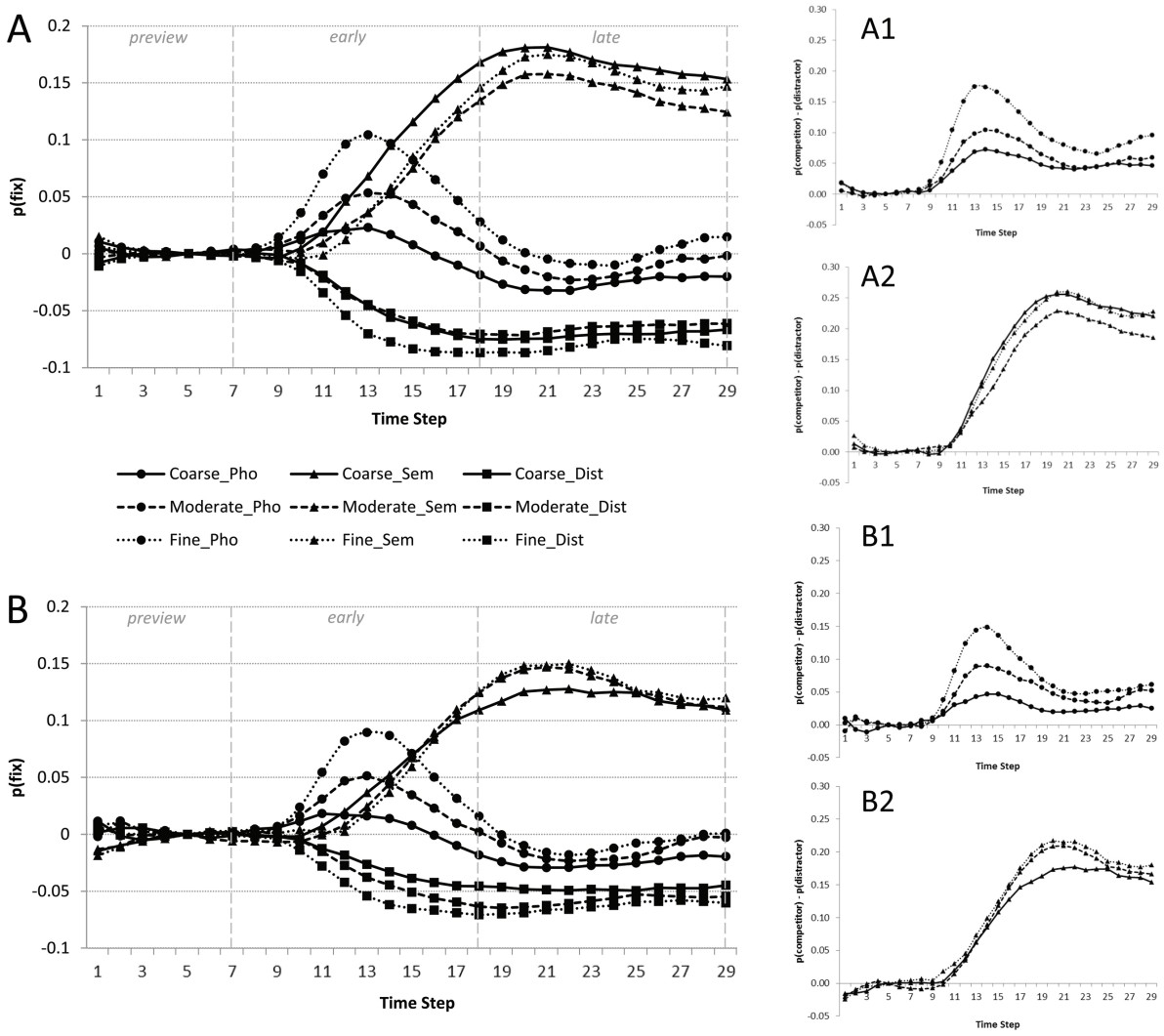

Figure 3: Time course of fixation behaviour displayed by the MIM model. Figures A and B display the proportion of fixations [p(fix)] directed towards items within scenes containing a phonological onset competitor (Pho), a semantic competitor (Sem) and two unrelated distractors. Fixation proportions are plotted for fine grain (Fine), moderate grain (Moderate) and coarse grain (Coarse) simulations with high cognitive efficiency (Figure 3A) and low cognitive efficiency (Figure 3B) (compare to behavioural data in Figures 1A and 1B). Figures A1 and B1 display the difference in fixation of phonological competitors compared to distractors in the high cognitive efficiency condition and low cognitive efficiency condition respectively, while figures A2 and B2 display the difference in fixation of semantic competitors compared to distractors in the high cognitive efficiency condition and low cognitive efficiency condition respectively.

For each time window in each test trial we calculated the total number of time steps a given category of item was fixated, *fix(item category)*. As two unrelated distractors were present in

each scene, for each time window (preview [time step 0-7], early [8-18], late [19-29], or early & late [8-29]) we divided the total number of fixations towards unrelated distractors by two. These totals were then used to calculate the empirical logits (log odds were used to avoid issues arising from calculating estimates based on proportion data, see Jaeger, 2008) for each category of item within each time window. We then calculated the difference between the log-odds of fixating a given competitor type and the log-odds of fixating unrelated distractors. This difference (*competitor bias*, see equation 1) formed our dependent measure as it reflects the difference in fixation behaviour as a consequence of representational overlap. Separate analyses were conducted for each competitor type (semantic competitor – unrelated distractor [semantic bias], and phonological competitor – unrelated distractor [phonological bias]).

Our initial analysis compared across simulations to examine whether the difference between fixation of competitor and distractor altered between the preview period and the period post target word onset (interest window = time steps 8 - 29), and whether this difference was influenced by either decreases in the granularity of phonological processing (i.e. whether behaviour in coarse and moderate grain size simulations differed from fine grain size simulations) or as a consequence of differences in cognitive efficiency. To compare behaviour between simulations we predicted the dependent measures with fixed effects: phonological grain size (coarse, moderate or fine: with fine grain mapped onto the intercept forming the baseline condition), cognitive efficiency (coded as a numerical factor centred on zero: high = -0.5, low = 0.5) and time window (coded as a numerical factor centred on zero: preview = -0.5, interest window = 0.5). The random effects structure included random intercepts for both instantiation and item as well as random slopes for time window both by instantiation and by item. This is the maximal random effect structure (Barr, Levy, Scheepers, & Tily, 2013), as both grain and noise were varied between instantiation and item. To derive p-values we assumed t-values were drawn from a normal distribution (Barr, 2008).

$$competitor\ bias = \log\left(\frac{fix(competitor)+0.5}{fix(competitor)-total\ time\ steps+0.5}\right) - \log\left(\frac{fix(distractor)+0.5}{fix(distractor)-total\ time\ steps+0.5}\right) \quad (1)$$

Results of this analysis showed that simulations displayed an increased *phonological bias* post word onset ($\beta = 0.454$, SE = 0.056, $t = 8.064$, $p<0.001$). Coarse grain simulations differed from fine grain simulations displaying a reduced *phonological bias* ($\beta = -0.202$, SE =

0.036, t = -5.551, p<0.001), importantly there was also an interaction between grain size (fine compared to coarse) and time window showing that fine grain simulations compared to coarse grain simulations displayed a further increased *phonological bias* post word onset (β = -0.338, SE = 0.073, t = -4.646, p<0.001). Fine grain simulations also displayed an increased *phonological bias* compared to moderate grain simulations but only in the period post word onset (β = -0.182, SE =0.073, t = -2.509, p = 0.012). There was no effect of cognitive efficiency on *phonological bias*.

Applying the same analysis to examine *semantic bias* showed that simulations displayed an increased *semantic bias* post word onset (β = 0.853, SE = 0.060, t = 14.242, p<0.001). There was no difference in *semantic bias* between moderate and fine grain simulations. The only significant difference between simulations of differing grain size was in the significant interaction for coarse compared to fine grain and cognitive efficiency (β = -0.167, SE = 0.074, t = -2.248, p = 0.025). This shows that reductions in cognitive efficiency lead to a lower *semantic bias* in coarse grain simulations compared to fine grain simulations. There was also an overall marginally significant increased *semantic bias* as a result of reductions in cognitive efficiency (β = 0.100, SE = 0.052, t = 1.905, p = 0.057).

We also applied the same approach to examine differences between simulations over the three theoretically motivated time regions (preview, early, and late), as previous research suggests that processing may differ in how effects are distributed across time, post word onset (e.g. Huettig & McQueen, 2007). This was done by splitting the remainder of the test trial, post preview, equally into two further windows, an early window (time step 8-18) and a late window (time step 19 – 29). The model structure used for analysis was identical to previous analyses except that we also included the additional window in the fixed effect of time window (preview, early or late: with preview mapped onto the intercept forming the baseline condition).

Results showed that simulations displayed a significant increase in *phonological bias* in the early versus preview window (β = 0.531, SE = 0.075, t = 7.048, p<0.001). This effect was only marginally significant in the late window (β = 0.148, SE = 0.085, t = 1.737, p = 0.082). A positive yet lower parameter estimate in late over early windows suggests a reduced bias towards fixating phonological competitors in later windows. We did not observe a main effect of either coarse or moderate grain. However, there was an interaction between coarse compared to fine grain and early versus preview window (β = -0.411, SE = 0.098, t = -4.202,

p<0.001), and coarse compared to fine grain and late versus preview window (β = -0.300, SE = 0.098, t = -3.063, p = 0.002). This means that although coarse and fine grain simulations did not differ in the preview period they did differ in both the early and late window. Parameter estimates indicate that the phonological effect was stronger in the fine grain simulation than the coarse grain simulation in both early and late windows, although a lower parameter estimate suggests this difference was lower in the late window. An interaction between moderate versus fine grain and early versus preview window (β = -0.230, SE = 0.098, t = -2.351, p = 0.019) was also observed, however there was no interaction between moderate versus fine grain and late versus preview window. This shows that the phonological effect only differed between moderate and fine grain simulations in the early window. A negative parameter estimate suggests fine grain simulations displayed a greater phonological effect in this window than moderate grain simulations. Thus, cognitive efficiency had no influence on *phonological bias*, with no significant main effect nor interactions.

Applying the same model structure to predict *semantic bias* yielded a main effect for both early (β = 0.499, SE = 0.073, t = 6.877, p<0.001) and late windows (β = 1.261, SE = 0.087, t = 14.475, p = 0.001). Increasing positive parameter estimates suggest an increasing bias toward fixating semantic competitors in early and late windows over preview periods, with a greater bias displayed in late over early windows. No other model parameters were significant.

We also examined the effects displayed by coarse, moderate and fine grain simulations individually. For each grain size we again used mixed effects models to predict measures of *semantic bias* and *phonological bias*. Models included fixed effects of time window (preview, early or late: with preview mapped onto the intercept forming the baseline condition) and cognitive efficiency (coded as a numerical factor centred on zero: high = -0.5, low = 0.5), in addition to a random effects structure that included random intercepts for both instantiation and item, and random slopes for time window both by instantiation and item (maximal random effects structure, Barr, Levy, Scheepers, and Tily, 2013).

Fine grain simulations displayed increased fixation of phonological competitors over unrelated distractors in the early window compared to the preview window (β = 0.531, SE = 0.075, t = 7.081, p<0.001), yet no difference was observed in this measure in the late window when compared to preview. Moderate grain simulations displayed a similar pattern of behaviour with an increased bias towards fixating phonological competitors in early windows

over preview (β = 0.301, SE = 0.071, t = 4.228, p<0.001), yet no difference between late windows and preview. Coarse grain simulations however, did not display a significant difference in *phonological bias* in either early or late windows when compared to preview. There was no evidence for an influence of cognitive efficiency on *phonological bias* for coarse, fine or moderate simulations.

Fine, coarse and moderate grain size simulations all displayed an increased bias towards fixating semantic competitors over distractors in both early and late windows when compared to the baseline preview window (fine [early: β = 0.499, SE = 0.076, t = 6.596, p<0.001; late: β = 1.261, SE = 0.106, t = 11.880, p<0.001], moderate [early: β = 0.450, SE = 0.073, t = 6.178, p<0.001; late: β = 1.196, SE = 0.103, t = 11.666, p<0.001]; coarse [early: β = 0.474, SE = 0.085, t = 5.586, p<0.001; late: β = 1.184, SE = 0.120, t = 9.865, p<0.001]). There was no evidence for an influence of cognitive efficiency on *semantic bias* in either the moderate or fine grain simulations. However, in the case of coarse grain simulations there was a significant interaction between cognitive efficiency and the late versus early time window (β = -0.386, SE = 0.138, t = -2.786, p = 0.005). A negative parameter estimate indicates that low cognitive efficiency leads to a reduction in the magnitude of the semantic effect in the late window over preview, for coarse grain simulations.

A further post-hoc analysis was also conducted to examine whether there was any evidence in the fixation behaviour of coarse grain simulations, post word onset, for an effect of phonological overlap. The phonological competitor bias in the preview period was compared to the same measure aggregated over all time steps in both early and late windows (cf. Huettig, Singh, & Mishra, 2011, Experiment 2). The model used for this analysis was identical to that used previously to examine effects at a single grain size, the only difference being to the fixed effect of time window (coded as a numerical factor centred on zero: preview = -0.5, combined early and late window = 0.5). This analysis showed that coarse grain simulations displayed a marginally significant increased bias in fixating phonological competitors over distractors post word onset (β = 0.116, SE = 0.060, t = 1.942, p = 0.052). There was no evidence from this analysis for an effect of cognitive efficiency (β = -0.052, t = -0.497, p = 0.619).

In summary, both moderate and fine grain simulations displayed a phonological effect limited to early periods post word onset, with fine grain simulations displaying an increased *phonological bias* over moderate grain simulations in this period. Coarse grain simulations in

contrast displayed a marginal increase in fixation of phonological competitors over unrelated distractors only when aggregating fixation across all time windows post word onset. Compared to fine grain simulations coarse grain simulations displayed a reduced *phonological bias* in both early and late windows post word onset. There was no evidence for an effect of cognitive efficiency on *phonological bias*.

Conversely, a marginal effect of cognitive efficiency was observed on *semantic bias* but only when pooling fixation behaviour across all time windows post word onset. All simulations displayed an increased *semantic bias* in early and late windows. The only difference between simulations of differing granularity in their *semantic bias* was observed in the interaction between coarse versus fine grain and cognitive efficiency. Also coarse grain simulations displayed a reduced *semantic bias* in late windows as a consequence of a reduction in cognitive efficiency.

4. Discussion

Our modelling results successfully replicated qualitative differences observed between high and low literates in sensitivity to phonological competitors, reported in Huettig, Singh and Mishra (2011). As was displayed in the behaviour of high literates, fixation of phonological onset competitors by fine and moderate grain size simulations was closely time locked to overlap between the competitor's phonological representation and the unfolding speech signal. Fine and moderate grain size simulations displayed an initial bias towards fixating phonological competitors shortly after word onset, and a rapid decline in fixation and return to baseline distractor levels once the phonological signal mismatched. Coarse grain simulations on the other hand displayed less dynamic changes in fixation of phonological competitors in response to overlap in the speech signal and fixated phonological competitors at levels close to unrelated distractors. Unlike fine and moderate simulations, for coarse grain simulations a marginal bias toward fixating phonological competitors over distractors was only observed when pooling fixation behaviour across both early and late windows. Such behaviour is similar to that displayed by low literates (Huettig, Singh & Mishra, 2011): When presented with scenes containing semantic and phonological competitors low literates did not display a bias towards phonological competitors. When tested under more sensitive conditions, in which only phonological competitors were present, low literates did display sensitivity to phonological overlap and looked marginally more towards phonological

competitors compared to distractors in the first 1000ms post word onset. As in the case of the coarse grain size simulations, low literates did not display the rapid increase and decrease in looks towards phonological competitors in response to signal overlap as was shown by high literates.

Our simulations therefore demonstrate that differences in the granularity of phonological processing can modulate the phonological effect displayed in studies of language mediated visual attention, reflecting effects of literacy in the use made of phonological information in processing visual and semantic representations of stimuli. This feature of model behaviour was largely driven by the fact that more fine grained representations more precisely encode the regions of the word's phonological representation that differ or overlap. This information can then be exploited by the system to dynamically adjust fixation behaviour.

In contrast to the phonological effects, grain size did not modulate the magnitude of the semantic effect. In Huettig, Singh and Mishra (2011), differences in fixation of semantic competitors between high and low literates were only observed in later time windows, once the spoken word has unfolded. Within our simulations, fixation of semantic competitors is dependent on the level of activation of semantic properties shared by the semantic competitor and target. This activation is maximal when the entire phonological representation of the target has been input to the model. At this point, at the word level, representational overlap does not differ across grain sizes and therefore as the simulations show we do not observe a difference in semantic competitor fixations as a result of grain size manipulations.

For the cognitive efficiency manipulation in the model, it was predicted that phonological effects would be less evident than semantic effects following reductions in cognitive efficiency, because within the model, activation of semantic representations is entirely dependent on the efficiency of information transfer within the network, unlike phonological representations which are provided as a direct input. Pilot simulations demonstrated that increasing noise levels within the network (i.e. $N[\mu=0, \sigma^2=0.05]$) were unlikely to lead to modulation of phonological effects, as this led to a failure in the model's ability to learn training tasks, tasks that we know both literates and illiterates are able to perform accurately. We also suggest that alternative methods of implementing cognitive efficiency within neural networks would also not simulate the pattern of the low-literate participants. For instance, reduction in the resources available for forming mappings between representations (Harm & Seidenberg, 1999) would impede the model's ability to learn tasks that both high and low

literacy groups are capable of performing, such as vision to semantics, or phonological to semantics mappings. A further alternative implementation by reducing the overall levels of activation passing between layers within the model (Monaghan & Shillcock, 2004), or by increasing the threshold such that more activation is required before activating a response, is also unlikely to simulate the focused distinctions between high and low literate populations. Such an implementation would result in the same information being processed, just requiring longer in order to be processed. Thus, the same peaks of performance for the phonological competitors condition would be observed, but at a later point in time, and similar delays for all other mapping tasks that the model is required to perform. Therefore, phonological effects would still be observed yet would be delayed, along with delays to a broad range of other tasks, which is not the behaviour displayed by low literates. Although we accept that it is possible for other implementations of reductions in cognitive efficiency to have implications for processing beyond those captured by our simulations, our results are sufficient to indicate that reducing the quality of information transfer within networks was not adequate for explaining the qualitative difference in sensitivity to phonological overlap displayed by low literates in Huettig, Singh and Mishra (2011).

Also replicating the behaviour of both low and high literates, all simulations displayed an increasing bias toward fixating semantic competitors, across early and late windows, when compared to preview periods. Further, similar to the quantitative difference observed between high and low literates, the semantic bias was greater for fine grain simulations compared to coarse grain simulations with low cognitive efficiency. Analysis of the influence of cognitive efficiency on coarse grain simulations shows that, unlike fine and moderate simulations, who displayed no cognitive efficiency effects, low cognitive efficiency lead to a reduction in semantic bias in late windows. This is similar to the observed behaviour of low literates, as it is only in late windows that levels of semantic bias are significantly lower than those displayed by high literates, with a lower asymptote in fixation of semantic competitors for low literates. Semantic bias is dependent on the activation of semantic properties shared by both target and competitor. The strength of this effect is dependent on the level to which these units are activated by the visual input provided by the semantic competitor, and the phonological input from the spoken target word. Reducing the efficiency of information transfer within the network will reduce the strength of the signal travelling from visual and phonological layers, to activate associated semantic layer units, and hence the level of activation of overlapping semantic features. Unlike fine and moderate simulations, low

cognitive efficiency only led to reduced semantic bias in coarse grain simulations. We suggest that this is most likely due to the componential structure of phonological representations in moderate and fine grain simulations that ensured activation of semantic features by phonological input was more robust to the introduction of noise. Importantly the differences between coarse and fine grain simulations captured the quantitative rather than qualitative differences in semantic bias observed between high and low literates. Within the model the effect of noise (cognitive efficiency) did not qualitatively alter the structure of information processed, but instead reduced the level to which overlapping semantic representations were activated and therefore semantic competitors fixated. Thus, it affected the role of representational overlap and the resulting fixation behaviour quantitatively rather than qualitatively.

Of the simulations conducted, only a comparison between a fine grain, high cognitive efficiency simulation and a coarse grain, low cognitive efficiency simulation replicated both the qualitative difference in sensitivity to phonological onset competitors and quantitative difference in sensitivity to semantic competitors observed between high and low literates in Huettig, Singh, and Mishra (2011). Therefore, our simulations suggest differences in language mediated visual attention as a consequence of literacy training may well be driven by both mechanisms: changes to phonological encoding as well as increased cognitive efficiency. The model thereby presents the first explicit account of the consequences of literacy that extend beyond phonological processing to other aspects of language processing and highlights the necessity for multimodal computational models in order to gain insight into the inherently complex issue of multimodal interaction within human cognitive processing.

Although there is substantial evidence for an effect of literacy on speech processing, there have been very few computational modelling studies that focus on understanding the emergent consequences of training on orthographic mappings for phonological or semantic systems involved in speech processing. For instance, previous models of reading acquisition have made important contributions demonstrating an influence of orthographic transparency on phonological processing (Harm & Seidenberg, 1999; Yang, McCandliss, Shu & Zevin, 2009) and semantic processing (Harm and Seidenberg, 2004; Yang, Shu, McCandliss & Zevin, 2013). However, such models have tended to be trained on prototypical phonological representations in which substantial phonological structure is embedded, and then the

processing of the phonological structure itself is investigated as both the input and output system. Such features of a model will have dramatic consequences for the type of structure to which the system develops sensitivity, and therefore these previous modelling studies are likely to have misrepresented the impact of orthographic training on the effects of the phonological grain size on processing within the language system more generally.

There remains a gap in our understanding of the extent to which literacy alters online speech processing (and broader aspects of cognitive processing) and the mechanisms through which it exerts an influence. To date the most influential cognitive models of speech processing do not allow for an influence of literacy on this process and have focused on modelling the behaviour of alphabetic literates (e.g. Distributed Cohort Model: Gaskell & Marslen-Wilson, 1997; TRACE: McClelland & Elman, 1986; Shortlist B: Norris & McQueen, 2008). We emphasize that a model of human speech processing should be sufficient to describe representations within the language system and their interaction, adequate for accounting for behaviour of literate and illiterate participants, as well as literates learning from different orthographies. The model presented in this paper does not simulate the emergent processes by which exposure to orthographic mappings leads to a restructuring of phonological representations or improved cognitive efficiency. It is possible that training on orthographic mappings also has emergent consequences for semantic processing that then give rise to increased sensitivity to semantic overlap. The empirical evidence suggests at most only a subtle effect of literacy on semantic processing (Da Silva et al., 2004; Kosmidis et al., 2004), however previous computational modelling studies of reading acquisition demonstrate that differences in the orthographic transparency of a language can have implications for the distribution of labour between phonological, semantic and orthographic processing networks (Harm and Seidenberg, 2004; Yang et al., 2013). Without knowing the emergent consequences of the additional orthographic mapping performed by such networks, it is not possible to rule out the possibility that such training could result in a modulation of semantic competitor effects without requiring a reduction in cognitive efficiency to be implemented.

It is important to emphasise that the MIM model was not specifically designed to simulate effects of literacy, but rather was an effective model of multimodal effects on language processing that was co-opted to extend to testing theories of literacy. However, a further means of validating the MIM model is to examine its additional predictions. Our results suggest that the representations governing language mediated eye gaze in low literates are

more coarse grained and therefore gaze is less sensitive to the temporal location of phonological overlap. One means of testing the coarseness of low literates' phonological representations would be through examining their sensitivity to phonological rhyme overlap. It has been previously observed that literate individuals display greater sensitivity to phonological overlap in the onset of words than in the rhyme (Allopenna et al., 1998). This may be because, in the case of rhyme competitors, by the time later overlapping phonemes unfold, earlier phonological information can be used to eliminate the rhyme competitor as a possible target, hence onset competitors are fixated more than rhyme competitors. If the model's coarse grain representations equate to word level representations in low literates then such representations would be unsuitable for determining whether the overlap occurs in the onset or rhyme of a given word. If onset and rhyme are matched for length, then the word level representations will be equally similar and therefore will generate fixations equally in a network processing at this word level. The model therefore predicts that unlike high literates, low literates should display little difference in their bias towards fixating phonological onset and phonological rhyme competitors.

A second prediction of the effects of orthographic transparency on language mediated visual attention also follows from the framework outlined in this paper. Our simulations indicated that training on orthographic mapping is critical to developing more fine grained processing of phonological information and subsequently displaying increased sensitivity to phonological competitors. The processing level model on which our hypothesis was based, psycholinguistic grain size theory, posits that the level of transparency between orthography and phonology determines the granularity of processing that is developed. It then follows that in non-alphabetic languages, in which there is little componentiality in the correspondence between the orthography and the speech sounds that make up a word, literacy training will have little effect on the granularity of speech processing. Therefore, logographic literates should behave more like illiterates on tasks that aim to measure this aspect of processing. There is already substantial empirical evidence within the literature to support such a position with Chinese literates who have not been exposed to an alphabetic writing system displaying reduced levels of phonological awareness (Cheung et al., 2001; Ho & Bryant, 1997; Huang & Hanley, 1995, 1997; McBride-Chang et al., 2004; Read et al., 1986; Shu, Peng & McBride-Chang, 2008). Further, recent evidence from neuroimaging studies support the critical role of orthographic transparency in modulating effects of literacy on speech processing, with less involvement of associated orthographic processing regions observed in logographic literates

compared to alphabetic literates when processing speech (Cao et al., 2011) and greater developmental changes in phonological processing regions as a consequence of literacy training in English over Chinese students (Brennan et al., 2013).

These results suggest that there may be similarities between illiterates and logographic literates in their phonological processing, and are compatible with the argument that literacy training on an alphabetic language leads to rearrangement in phonological processing networks such that phonological processing becomes more fine grained. Should the phonological effect observed in studies of language mediated eye gaze be modulated by the granularity of speech processing in the manner our simulations suggest, then we would predict that logographic literates (not exposed to training on alphabetic systems e.g. Hanyu Pinyin) like illiterates should display reduced sensitivity to phonological overlap compared to alphabetic literates, and, unlike alphabetic literates, their fixation of such competitors should not be tightly time locked to overlap in the speech signal. However if, as our simulations indicate, quantitative differences in the semantic bias observed between high and low literates result from increased efficiency of information transfer within the networks trained during literacy acquisition, we would not predict a reduction in semantic bias in logographic literate populations, as training of relevant networks should be similar in both literate groups.

To conclude, influential models of human language processing have been developed largely only with reference to the behaviour of alphabetic literates, and generally do not take into account the influence of literacy, or of varying orthographic systems, on the processing system. Those that currently do are likely to underrepresent its consequences because of their inclusion of pre-specified componential phonological representations. Our modelling work using the MIM model demonstrated that two competing theories of effects of literacy on language learning may well be compatible and complementary contributors to language processing: both cognitive efficiency and phonological grain-size differences were required to simulate the detailed data on phonological and semantic processing in literate and illiterate participants. Given that approximately 16% of the adult human population are illiterate (UNESCO Institute for Statistics, 2013) and a further 15%[2] (approximately) of the human population are literate in logographic languages, understanding the consequences of these

[2] *2 Based on a world population of 7.148 billion (United States Census Bureau, World Population Clock, 2014) and a population of the Peoples' Republic of China aged over 15 years 1.108 billion (The World Bank, 2012) of which 94% are literate (UNESCO Institute for Statistics, 2012).*

factors for human cognition remains an important challenge for future research, for which multimodal computational models are likely to provide an informative, even necessary, tool.

References

Adrián, J. A., Alegria, J., & Morais, J. (1995). Metaphonological abilities of Spanish illiterate adults. *International Journal of Psychology, 30(3),* 329-351.

Alcock, K. J., Ngorosho, D., Deus, C., & Jukes, M. C. H. (2010). We don't have language at our house: disentangling the relationship between phonological awareness, schooling, and literacy. *British Journal of Educational Psychology, 80(1),* 55-76.

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science, 14(5),* 255-259.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language, 38(4),* 419-439.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59(4),* 390-412.

Barr, D. J. (2008). Analyzing 'visual world'eyetracking data using multilevel logistic regression. *Journal of memory and language, 59(4),* 457-474.

Bengtsson, S. L., Nagy, Z., Skare, S., Forsman, L., Forssberg, H., & Ullén, F. (2005). Extensive piano practicing has regionally specific effects on white matter development. *Nature neuroscience, 8(9),* 1148-1150.

Bramao, I., Mendonca, A., Faisca, L., Ingvar, M., Petersson, K. M., & Reis, A. (2007). The impact of reading and writing skills on a visuo-motor integration task: A comparison between illiterate and literate subjects. *Journal of the International Neuropsychological Society, 13(2),* 359-364.

Brennan, C., Cao, F., Pedroarena-Leal, N., McNorgan, C., & Booth, J. R. (2013). Reading acquisition reorganizes the phonological awareness network only in alphabetic writing systems. *Human brain mapping, 34(12),* 3354-3368.

Burnham, D. (2003). Language specific speech perception and the onset of reading. *Reading and Writing, 16(6),* 573-609.

Cao, F., Khalid, K., Lee, R., Brennan, C., Yang, Y., Li, K., Bolger, D. J. & Booth, J. R. (2011). Development of brain networks involved in spoken word processing of Mandarin Chinese. *NeuroImage, 57(3),* 750-759.

Caravolas, M., Volin, J., & Hulme, C. (2005). Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: Evidence from Czech and English children. *Journal of Experimental Child Psychology*, *92*, 107–139.

Chéreau, C., Gaskell, M. G., & Dumay, N. (2007). Reading spoken words: Orthographic effects in auditory priming. *Cognition, 102(3),* 341-360.

Cheung, H., Chen, H. C., Lai, C. Y., Wong, O. C., & Hills, M. (2001). The development of phonological awareness: Effects of spoken language experience and orthography. *Cognition, 81(3),* 227-241.

da Silva, C. G., Petersson, K. M., Faísca, L., Ingvar, M., & Reis, A. (2004). The effects of literacy and education on the quantitative and qualitative aspects of semantic verbal fluency. *Journal of Clinical and Experimental Neuropsychology, 26(2),* 266-277.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic bulletin & review, 12(3),* 453-459.

de Jong, P. F., & van der Leij, A. (2003). Developmental changes in the manifestation of a phonological deficit in dyslexic children learning to read a regular orthography. *Journal of Educational Psychology, 95(1),* 22-40.

Deary, I. J., Bastin, M. E., Pattie, A., Clayden, J. D., Whalley, L. J., Starr, J. M., & Wardlaw, J. M. (2006). White matter integrity and cognition in childhood and old age. *Neurology, 66(4),* 505-512.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J. & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science, 330(6009),* 1359-1364.

Dijkstra, T., Roelofs, A., & Fieuws, S. (1995). Orthographic effects on phoneme monitoring. *Canadian Journal of Experimental Psychology, 49(2),* 264-271.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience, 2(10),* 704-716.

Fields, R. D. (2008). White matter in learning, cognition and psychiatric disorders. *Trends in Neurosciences, 31(7),* 361-370.

Goswami, U. (2003). Why theories about developmental dyslexia require developmental designs. *Trends in Cognitive Sciences, 7(12),* 534-540.

Gutiérrez, R., Boison, D., Heinemann, U., & Stoffel, W. (1995). Decompaction of CNS myelin leads to a reduction of the conduction velocity of action potentials in optic nerve. *Neuroscience Letters, 195(2),* 93-96.

Garlock, V. M., Walley, A. C., & Metsala, J. L. (2001). Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language, 45(3),* 468-492.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review, 106(3),* 491-528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review, 111(3),* 662.

Ho, C. S. H., & Bryant, P. (1997). Phonological skills are important in learning to read Chinese. *Developmental Psychology, 33(6),* 946.

Hoonhorst, I., Medina, V., Colin, C., Markessis, E., Radeau, M., Deltenre, P., & Serniclaes, W. (2011). Categorical perception of voicing, colors and facial expressions: A developmental study. *Speech communication, 53(3),* 417-430.

Shu, H., Peng, H., & McBride-Chang, C. (2008). Phonological awareness in young Chinese children. *Developmental Science, 11(1),* 171-181.

Huang, H. S., & Hanley, J. R. (1995). Phonological awareness and visual skills in learning to read Chinese and English. *Cognition, 54(1),* 73-98.

Huang, H. S., & Hanley, J. R. (1997). A longitudinal study of phonological awareness, visual skills, and Chinese reading acquisition among first-graders in Taiwan. *International Journal of Behavioral Development, 20(2),* 249-268.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57(4),* 460-482.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137,* 151-171.

Huettig, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica, 121(1),* 65-80.

Huettig, F., & Altmann, G. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96(1),* B23-B32.

Huettig, F., & Altmann, G. T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition, 15(8),* 985-1018.

Huettig, F., Singh, N., & Mishra, R. K. (2011). Language-mediated visual orienting behavior in low and high literates. *Frontiers in Psychology, 2,* 285.

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The Causal Role of Phoneme Awareness and Letter-Sound Knowledge in Learning to Read Combining Intervention Studies With Mediation Analyses. *Psychological Science, 23(6),* 572-577.

Hulme, C., Snowling, M., Caravolas, M., & Carroll, J. (2005). Phonological skills are (probably) one cause of success in learning to read: A comment on Castles and Coltheart. *Scientific Studies of Reading, 9(4),* 351-365.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language, 59(4),* 434-446.

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta psychologica, 86(2),* 199-225.

Kello, C. T., & Plaut, D. C. (2000). Strategic control in word reading: Evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(3),* 719.

Kolinsky, R., Cary, L., & Morais, J. (1987). Awareness of words as phonological entities: The role of literacy. *Applied Psycholinguistics, 8(3),* 223-232.

Kolinsky, R., Pattamadilok, C., & Morais, J. (2012). The impact of orthographic knowledge on speech processing. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies, (63),* 161-186.

Kosmidis, M. H., Tsapkini, K., Folia, V.., Vlahou, C. H., & Kiosseoglou, G.. (2004). Semantic and phonological processing in illiteracy. *Journal of the International Neuropsychological Society, 10(6),* 818-827.

Kraft, R. H., Mitchell, O. R., Languis, M. L., & Wheatley, G. H. (1980). Hemispheric asymmetries during six-to eight-year-olds performance of Piagetian conservation and reading tasks. *Neuropsychologia, 18(6),* 637-643.

Levi-Bruhl, (1923). *Primitive mentality. Translated from the French by Lilian A. Clare.* London: George Allen & Unwin. (Orig. pub. Paris, 1922).

Li, S. C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science, 15(3),* 155-163.

Li, Y., Liu, Y., Li, J., Qin, W., Li, K., Yu, C., & Jiang, T. (2009). Brain anatomical network and intelligence. *PLoS Computational Biology, 5(5),* e1000395.

Loureiro, C. D. S., Willadino Braga, L., Souza, L. D. N., Queiroz, E., & Dellatolas, G. (2004). Degree of illiteracy and phonological and metaphonological skills in unschooled adults. *Brain and language, 89(3),* 499-502.

Luria, A. (1976). *Cognitive development: Its cultural and social foundations.* Cambridge: Cambridge University Press.

Madden, D. J., Bennett, I. J., & Song, A. W. (2009). Cerebral white matter integrity and cognitive aging: contributions from diffusion tensor imaging. *Neuropsychology review, 19(4),* 415-435.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8(1),* 1-71.

McBride-Chang, C., Bialystok, E., Chong, K. K., & Li, Y. (2004). Levels of phonological awareness in three cultures. *Journal of Experimental Child Psychology, 89(2),* 93-111.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology, 18(1),* 1-86.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & cognition, 37(7),* 1026-1039.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of memory and language, 59(4),* 475-494.

Monaghan, P., & Shillcock, R.C. (2004). Hemispheric asymmetries in cognitive modelling: Connectionist modelling of unilateral visual neglect. *Psychological Review, 111*, 283-308.

Morais, J., Bertelson, P., Cary, L., & Alegria, J. (1986). Literacy training and speech segmentation. *Cognition, 24(1),* 45-64.

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition, 7(4),* 323-331.

Morrison, F. J., Smith, L., & Dow-Ehrensberger, M. (1995). Education and cognitive development: A natural experiment. *Developmental Psychology, 31(5),* 789.

Morais, J., Content, A., Cary, L., Mehler, J., & Segui, J. (1989). Syllabic segmentation and literacy. *Language and Cognitive Processes, 4(1),* 57-67.

Muneaux, M., & Ziegler, J. (2004). Locus of orthographic effects in spoken word recognition: Novel insights from the neighbour generation task. *Language and Cognitive Processes, 19(5),* 641-660.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review, 115(2),* 357-395.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23(03),* 299-325.

Olivers, C. N. L., Huettig, F., Singh, J. P., & Mishra, R. K. (in press). The influence of literacy on visual search. *Visual Cognition.*

Pattamadilok, C., Knierim, I. N., Duncan, K. J. K., & Devlin, J. T. (2010). How does learning to read affect speech perception? *The Journal of Neuroscience, 30(25),* 8435-8444.

Pattamadilok, C., Perre, L., Dufau, S., & Ziegler, J. C. (2009). On-line orthographic influences on spoken language in a semantic task. *Journal of Cognitive Neuroscience, 21(1),* 169-179.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. Neural Computation, 1(2), 263–269.

Penke, L., Maniega, S. M., Murray, C., Gow, A. J., Hernández, M. C. V., Clayden, J. D., Starr, J. M., Wardlaw, J. M., Bastin, M. E. & Deary, I. J. (2010). A general factor of brain white matter integrity predicts information processing speed in healthy older people. *The Journal of Neuroscience, 30(22),* 7569-7574.

Perre, L., & Ziegler, J. C. (2008). On-line activation of orthography in spoken word recognition. *Brain Research, 1188,* 132-138.

Perre, L., Midgley, K., & Ziegler, J. C. (2009). When beef primes reef more than leaf: orthographic information affects phonological priming in spoken word recognition. *Psychophysiology, 46(4),* 739-746.

Perre, L., Pattamadilok, C., Montant, M., & Ziegler, J. C. (2009). Orthographic effects in spoken language: On-line activation or phonological restructuring?. *Brain research, 1275,* 73-80.

Pujol, J., Soriano-Mas, C., Ortiz, H., Sebastian-Galles, N., Losilla, J. M., & Deus, J. (2006). Myelination of language-related areas in the developing brain. *Neurology, 66(3),* 339-343.

R Development Core Team. (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Read, C., Yun-Fei, Z., Hong-Yin, N., & Bao-Qing, D. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition, 24(1),* 31-44.

Reis, A., & Castro-Caldas, A. (1997). Illiteracy: A cause for biased cognitive development. *Journal of the International Neuropsychological Society, 3(05),* 444-450.

Reis, A., Guerreiro, M. & Petersson, K. M. (2003). A sociodemographic and neuropsychological characterization of an illiterate population. *Applied Neuropsychology, 10(4),* 191–204.

Salthouse, T. A. (1988). Initiating the formalization of theories of cognitive aging. *Psychology and aging, 3(1),* 3-16.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological review, 103(3),* 403-428.

Salthouse, T. A. (2005). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology, 19(4),* 532-545.

Scliar-Cabral, L., Morais, J., Nepomuceno, L. & Kolinsky, R. (1997). The awareness of phonemes: So close-so far away. *International Journal of Psycholinguistics, 13,* 211-240.

Serniclaes, W., Ventura, P., Morais, J., & Kolinsky, R. (2005). Categorical perception of speech sounds in illiterate adults. *Cognition, 98(2),* B35-B44.

Shu, H., Peng, H., & McBride-Chang, C. (2008). Phonological awareness in young Chinese children. *Developmental Science, 11(1),* 171-181.

Smith, A. C., Monaghan, P., & Huettig, F. (2013a). An amodal shared resource model of language-mediated visual attention. *Frontiers in Psychology, 4.*

Smith, A.C., Monaghan, P., & Huettig, F. (2013b). Modelling language-vision interactions in the hub-and-spoke framework. In J. Mayor, & P. Gomez (Eds.), *Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop (NCPW13).* Singapore: World Scientific Publishing.

Storkel, H. L. (2002). Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language, 29(2),* 251-274.

Stoodley, C. J., & Stein, J. F. (2006). A processing speed deficit in dyslexic adults? Evidence from a peg-moving task. *Neuroscience letters, 399(3),* 264-267.

Szwed, M., Ventura, P., Querido, L., Cohen, L., & Dehaene, S. (2012). Reading acquisition enhances an early visual process of contour integration. *Developmental science, 15(1),* 139-149.

Taft, M. (2006). Orthographically influenced abstract phonological representation: Evidence from non-rhotic speakers. *Journal of psycholinguistic research, 35(1),* 67-78.

Taft, M., & Hambly, G. (1985). The influence of orthography on phonological representations in the lexicon. *Journal of Memory and Language, 24(3),* 320-335.

The World Bank (2014, Feb 21). Development Indicators. Retrieved February, 21, 2014, from http://data.worldbank.org/indicator/SP.POP.0014.TO.ZS/countries

Tolhurst, D. J., & Lewis, P. R. (1992). Effect of myelination on the conduction velocity of optic nerve fibres. *Ophthalmic and Physiological Optics, 12(2),* 241-243.

Treiman, R., & Zukowski, A. (1991). Levels of phonological awareness. In S. A. Brady, & D. P. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman,* (pp. 67-83). Oxford: Routledge.

Turken, A. U., Whitfield-Gabrieli, S., Bammer, R., Baldo, J. V., Dronkers, N. F., & Gabrieli, J. D. (2008). Cognitive processing speed and the structure of white matter pathways: convergent evidence from normal variation and lesion studies. *Neuroimage, 42(2),* 1032-1044.

United States Census Bureau. (2014, Feb 21). World Population Clock. Retrieved February, 21, 2014, from http://www.census.gov/popclock/

UNESCO Institite for Statistics. (2012). Adult and Youth Literacy, 1990-2015, Montreal: UNESCO.

UNESCO Institute for Statistics. (2013). Adult and Youth Literacy Fact Sheet, Montreal: UNESCO.

Ventura, P., Kolinsky, R., Querido, J. L., Fernandes, S., & Morais, J. (2007). Is phonological encoding in naming influenced by literacy?. *Journal of psycholinguistic research, 36(5),* 341-360.

Ventura, P., Morais, J., Pattamadilok, C., & Kolinsky, R. (2004). The locus of the orthographic consistency effect in auditory word recognition. *Language and Cognitive Processes, 19(1),* 57-95.

Vygotsky, L. S.(1978). *Mind in society: The development of higher psychological processes society* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.), Harvard University Press, Cambridge.

Waxman, S. G. (1980). Determinants of conduction velocity in myelinated nerve fibers. *Muscle & nerve, 3(2),* 141-150.

Yang, J., Shu, H., McCandliss, B. D. & Zevin, J. D. (2013). Orthographic influences on division of labor in learning to read Chinese and English: Insights from computational modeling. *Bilingualism: Language and Cognition, 16(2),* 354-366.

Yang, J., McCandliss, B. D., Shu, H., & Zevin, J. D. (2009). Simulating language-specific and language-general effects in a statistical learning model of Chinese reading. *Journal of memory and language, 61(2),* 238-257.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(1),* 1-14.

Yu, C. & Ballard, D.H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70*, 2149-2165.

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading a cross-language investigation. *Psychological Science*, *21,* 551-559.

Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review, 5(4),* 683-689.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin, 131(1),* 3-29.

Appendix

Neural networks simulations were conducted using Mikenet version 8.0 developed by M. W. Harm (www.cnbc.cmu.edu/~mharm/research/tools/mikenet/), a collection of libraries written in the C programming language for implementing and training connectionist networks.

Networks were trained using the continuous recurrent backpropagation through time training algorithm provided in Mikenet (crbp.c) which implements Pearlmutter (1989). Unit activation was calculated using a logistic activation function and sum squared error was used to calculate error. Time within the network was modelled using 14 samples and an integration constant of 0.25. All other parameters were set to the default values implemented in Mikenet version 8.0.

Mixed effects model analysis was performed using the R (version 3.1.0; R Development Core Team, 2009) libraries lme4 (version 1.1-6) and languageR (version 1.4.1).

Chapter 6

The effects of orthographic transparency on the reading system: Insights from a computational model of reading development[1]

Abstract

Orthographic systems vary dramatically in the extent to which they encode a language's phonological and lexico-semantic structure. Studies of the effects of orthographic transparency suggest that such variation is likely to have major implications for how the reading system operates. However, such studies have been unable to examine in isolation the contributory effect of transparency due to co-varying linguistic or socio-cultural factors. Within the current study we train a connectionist implementation of the triangle model of reading on a range of orthographic systems representing the range of the world's writing systems (alphabetic; alphasyllabic; consonantal; syllabic; logographic) whilst controlling for phonological and semantic structure. We show that the triangle model is effective as a universal model of reading, able to replicate key behavioural and neuroscientific results, and generating new predictions deriving from an explicit description of the effects of orthographic transparency on how reading is realised.

[1] *Adapted from Smith, A. C., Monaghan, P., & Huettig, F. (in preparation). The effects of orthographic transparency on the reading system: Insights from a computational model of reading development.*

1. Introduction

Current dominant psychological and cognitive neuroscientific descriptions of how we read, how we acquire this ability and its broader cognitive consequences are almost entirely built upon the study of alphabetic literates, where readers transform series of letters each of which correspond, more or less, to phonological forms of words. The world's major orthographic systems (Comrie, 2013: alphabetic, consonantal, syllabic, alphasyllabic, logographic) vary dramatically in the manner in which they encode a language's phonological and semantic structure (i.e. their semantic or phonological transparency). Recent years have seen an increasing diffusion of attention within reading research to alternate orthographic systems. Across the world's literate population, there are over 1 billion logographic literates, over 500 million alphasyllabic literates in addition to many hundreds of million consonantal and syllabic literates. However, given this variety of orthographic systems, existing reading models are largely based on reading in English or other alphabetical orthographies (Coltheart et al., 2001; Harm & Seidenberg, 2004; Perry et al., 2010), yet the variety of ways in which writing systems reflect representations of words are likely to have profound consequences for how reading is acquired and the effect that orthography has on the realised cognitive mechanisms recruited for reading.

In this paper we present a series of models of reading that implement the inherent differences between writing systems across the world's major orthographic systems. We first provide a review of the literature that describes the effect that orthographic variation has on the trajectory of reading acquisition, the impact of orthographic systems on the processes involved in the mature reading system, and the differential effects of literacy on cognitive processing more broadly. We then present our implemented model of reading, based on the triangle modelling tradition in simulating the reading process (e.g., Harm & Seidenberg 2004), and demonstrate how the variation found in the world's orthographic systems affects both the manner in which reading is acquired and how the reading system operates after extended experience of reading. The modelling demonstrates that a comprehensive understanding of the acquisition and operation of reading requires a full consideration of variation in orthographies. Our modelling enables an explicit test of theoretical views on how orthographic variation affects the reading system, and provides an explanation for how behavioural distinctions in reading development emerge as a consequence of these variations in the way in which sound and meaning distinctions are formed in the writing system.

Orthographic Diversity

Five categories of orthographic system are typically defined to describe the range of extant writing systems found globally: alphabetic, consonantal, syllabic, alphasyllabic and logographic (Comrie, 2013). The extent to which the written form reflects the phonology of the word, the transparency of the orthography, varies greatly over these systems, and we describe them in order of transparency, from greatest to least.

Alphabetic systems are the dominant system in the 21st Century world existing throughout Europe, the Americas, Australasia and significant portions of Asia and sub-Saharan Africa. Within such systems the basic unit of representation is the phoneme, therefore the orthography contains detailed information regarding the fine grained phonological structure of the language. Alphabetic systems vary in the granularity and regularity of mappings between orthography and phonology. Within shallow alphabetic systems, such as Finnish or Serbo-Croatian, there is close to perfect one to one correspondence between individual phonemes and graphemes (grapheme: a letter or set of letters corresponding to a phoneme), whereas in deep alphabetic systems such as English there are deviations in regularity (that is, the extent to which a letter or set of letters maps onto the same phoneme or set of phonemes, e.g., fl<u>i</u>nt, p<u>i</u>nt) and granularity (that is, the number of letters that combine to relate to sounds in the word, e.g. c<u>o</u>t, y<u>ach</u>t) of mappings.

Consonantal systems, in contrast, possess very similar structural properties to alphabetic systems yet with the defining feature of representing only consonants, not vowels. Such systems (e.g. Arabic, Hebrew) are prevalent in the Middle East and northern Africa. Consonantal systems were in the past more widespread however many of today's consonantal systems also exist in alphabetic or alphasyllabic form due to the addition of diacritics that are used to indicate the presence of a particular vowel.

The basic grapheme within alphasyllabic systems indicates a consonant, however information regarding the vowel is also encoded in such systems either in the form of diacritics added to the consonant preceding or following the vowel or through a predefined transformation of the preceding or subsequent consonant's representation as a grapheme. Such systems are widespread concentrated most intensely in India (Devaganari, e.g. Hindi) and South East Asia (e.g. Thai).

Syllabic systems provide a fourth category, within which the functional unit is the syllable. In its idealized form a single grapheme corresponds to each syllable within the phonology.

Japanese hiragana is an example of such a system however pure syllabic systems are rare with Japanese today largely communicated in a mixed logographic-syllabic form.

The basic representational unit in logographic systems is the morpheme, and therefore contrasts with other orthographic systems in which representational units are related to phonological properties. Chinese is the only logographic script in wide spread use today. One further characteristic of Chinese characters is that most (82%; Zhou, 1978) contain phonetic radicals and semantic radicals that, respectively, provide probabilistic information regarding the word's phonetic or semantic properties.

Effects of orthographic transparency

In all these orthographic types, readers can learn to map from written to spoken and meaning representations for words. However, this diversity in orthographic structure can have quantitative and qualitative effects on acquisition and processing of the reading system, and may also have wider cognitive implications for the way in which the reading system integrates with pre-existing language processing networks.

Effects on acquisition

Probably the domain in which there is greatest understanding and consensus regarding the impact of orthographic transparency on reading is in the rate of acquisition of phonological decoding abilities (Snowling & Hulme, 2005). A consistent problem faced by researchers that aim to compare the effects of orthographic structure across contrasting systems is that systems and populations also vary across many other dimensions, such as language factors, e.g., phonological complexity of the language, visual complexity of the orthography, word order, morphological complexity, and syntactic structure; or socio-cultural factors, e.g., teaching methods, educational background, and student motivation. Nevertheless, there have been attempts to characterise differences that are a consequence only of the orthography (Seidenberg, 2013). It has been found that children learning a deep alphabetic system such as English require 4-5 years of literacy training in order to reach 90% accuracy on non-word reading tasks (Goswami, Gombert, & De Barrera, 1998) whereas children learning to read a shallow system such as Finnish reach this level of attainment after their first year of tuition (Seymour, Aro, & Erskine, 2003). Although it is not possible to control for factors beyond the orthographic system variation, studies that aim to minimise the impact of such factors have consistently shown that increased phonological transparency coincides with increased phonological decoding acquisition rates. For example, Genesee and Caravolas (1997)

compared groups of English speaking and French speaking children from the same region of Canada on monosyllabic non-word and word reading performance following a year of literacy tuition. They found that the English speaking population displayed 24% lower performance on word reading tasks and 27% lower performance on non-word reading tasks compared to their French speaking counterparts, which was suggested to be due to the greater orthographic transparency of French than English. A similar study conducted between children from the same region of the United Kingdom learning to read either Welsh or English found that children learning Welsh, the shallower orthographic system, similarly displayed increased performance on non-word and word reading when controlling for training exposure (Hanley, Masterson, Spencer, & Evans, 2004).

Direct comparisons across orthographic types are less common and prone to increased confounds of linguistic and socio-cultural factors, however the time required to reach similar levels of reading proficiency across populations is highly suggestive of transparency leading to faster rates of decoding acquisition. Nag (2007) demonstrated that a population learning to read Kannada, an alphasyllabic system, displayed delayed decoding abilities in comparison to English populations when exposure and quality of tuition was controlled for. Asfaha et al. (2009) compared reading acquisition rates over the first year of literacy acquisition across four populations learning to read one of four African languages in either a syllabic (Ge'ez) or alphabetic (Latin) script, and observed an increased rate of acquisition of the syllabic orthographies, which contrasts somewhat with other studies on effects of orthographic transparency. However, for acquisition of a logographic system, the results are again consistent with transparency affecting learning to read. Chinese was found to result in a slow acquisition rate as it requires intensive training over the first 6 years of schooling in order for children to learn the 2,500 foundational characters required to support proficient Chinese reading (Cheung & Ng, 2003).

Studies of reading acquisition that have compared across orthographies predictors of word reading that extend beyond phonological decoding (e.g. reading fluency) demonstrate variation in the influence of phonological, semantic and morphological factors. Ziegler et al., (2010) showed that across alphabetic systems of varying orthographic depth predictors of reading performance were largely consistent, yet varied systematically in their relative influence as a function of orthographic transparency. Specifically, phonological awareness was a stronger predictor in less transparent alphabetic scripts. By contrast results presented in

Cohen-Mimran, (2009) show that phonological awareness did not predict reading fluency in less transparent orthographies such as pointed (alphasyllabic) or unpointed (consonantal) Hebrew scripts, whereas morphological measures were a good predictors for both and semantic measures for unpointed performance. Similarly, Tong & McBride-Chang, (2010) tested reading performance of children learning to read in Chinese and English concurrently. Their results show that predictors of variation in reading performance in Chinese and English was stable across age groups yet differed across scripts, with morphological measures predicting variation in Chinese reading, yet not English while phonological awareness predicted reading in English but not Chinese (see also Tong, et al., 2009). Together this evidence argues for contrasting influences of semantic and phonological knowledge on reading acquisition as a consequence of orthographic transparency.

Although, many studies that examine the effects of orthographic transparency on reading acquisition report a delay in decoding ability in less transparent systems, the extent to which transparency impacts on comprehension skills is less clear as comprehension measures are often not included in such studies (Seidenberg, 2013). For example, in Turkish, a shallow orthographic system, a high proficiency in decoding is achieved very early however comprehension ability is delayed (Durgunglu, 2006). Similarly English speaking children have been shown to regularly understand the meaning of written words they are unable to decode accurately (Nation & Cocksey, 2009). This is potentially reflected in the Welsh–English study previously described which also found that English readers outperformed Welsh readers in their comprehension abilities (Hanley, Masterson, Spencer & Evans, 2004).

In summary, the data regarding effects of orthographic transparency on the ability to learn orthographic to phonological mappings generally demonstrates that transparency aids acquisition. However, the effect of transparency on comprehension remains an underexplored issue.

Effects on processing

Learning to read requires acquiring mappings from orthography onto a pre-existing system of phonological and semantic representations, derived through spoken language exposure. There are a variety of ways in which orthography can integrate with this existing phonology-semantics system, and the processing of the reading system may be profoundly constrained by the properties of the orthographic system being acquired.

There are two theoretically-motivated paths via which orthographic information could activate semantic and phonological representations, which are consistent with most implemented models of word reading. Activations can be either direct, where correspondences between orthographic and phonological forms and between orthographic and semantics are acquired with resources dedicated to forming these mappings. Alternatively, activations could be indirect, where correspondences between orthography and phonology are mediated by the word's meaning, where phonology to meaning representations are acquired prior to literacy, or from orthography to semantics, via phonological representations. The orthographic depth hypothesis [ODH] (Frost, Katz & Bentin, 1987; Katz & Feldman, 1981) states that the transparency of the orthography will dictate the extent to which direct and indirect paths are engaged for reading, and that this will be determined by the extent of the systematicity between orthographic and phonological or semantic representations.

The strongest interpretation of the ODH contends that literates of shallow alphabetic systems will rapidly acquire word naming fluency – the orthographic to phonological mappings – along a direct route because of the regularity of the grapheme to phoneme correspondences. However, reading comprehension – so mapping from orthography to semantics – will largely depend on the indirect route via phonology, because orthography to semantics is a largely arbitrary mapping, which is hard to learn, and so activation of meaning will derive from the systematic orthography to phonology combining with the pre-trained phonology to semantics system. Equally, a strong ODH position also argues that literates of opaque orthographies such as logographic systems will depend on a direct route from orthography to semantics as the complexities of the orthographic to phonological mappings mean that learning such direct mappings no longer provides an advantage. In slight contrast, the triangle model of reading (Seidenberg & McClelland, 1989), which implements interactive routes to and from orthography, phonology and semantic representations, suggests that both direct and indirect routes are likely to be actively recruited during reading but to differing degrees depending on the systematicity of the mapping from orthography to phonology and to semantics (Plaut et al., 1996). The triangle model is thus consistent with a weak ODH, which predicts that greater orthography to phonology transparency will result in greater distribution of labour across alternative paths for word meaning tasks, and that orthography to semantic transparency would result in increased distribution of labour across paths for word naming tasks (Harm & Seidenberg, 2004).

Computational models have been developed to investigate division of labour within the reading system, but these have largely been limited to alphabetic systems (Harm & Seidenberg, 2004; though see Yang, Shu, McCandliss, & Zevin, 2013). Harm & Seidenberg (2004) trained a connectionist computational implementation of the triangle model to learn mappings between orthographic, semantic and phonological representations. By lesioning paths within the model, they observed that at earlier stages of training an advantage was observed for word comprehension tasks by processing via the indirect phonological pathway (so from orthography to phonology to semantics, compared to from orthography directly to semantics). However, this advantage reduced over the course of training such that some words could be processed only by the direct orthography to semantic route, and by the end of training approximately half the corpus could be read by either route. Thus, this computational study suggests that even for alphabetic (but deep) systems such as English both routes are likely to be recruited.

The question remains however, is the triangle model adequate as a framework to explain reading development regardless of the orthographic system, and if so what does it reveal about how orthographic systems affect processing within the reading system? The computational modelling literature is divided on this issue, with distinct architectures devised for alphabetic and logographic systems (Perfetti et al., 2005). Proponents of the dual route cascaded model (DRC), initially developed for processing the deep alphabetic system of English, for example, have suggested that due to the scale of structural differences between alphabetic, syllabic and logographic systems the architecture they propose to support reading in alphabetic systems would not be applicable for syllabic or logographic systems (Coltheart et al., 2001). Recent computational modelling studies conducted by Yang et al. (2006; 2009; 2013) have made a substantial contribution to our understanding of the viability of the triangle model architecture to support Chinese reading and the effects of such an orthographic structure on processing. Yang et al. (2009) focused on the emergent properties of networks trained purely on direct orthographic to phonological Chinese mappings. They observed that this single path is able to develop internal representations that take advantage of the small degree of systematicity carried in the logographic orthography regarding the phonological properties of the word. Furthermore, Yang et al. (2006; 2013a) constructed a computational implementation of the triangle model similar to that of Harm & Seidenberg (2004) which was found to be able to support Chinese reading, and that, in comparison to English, it displayed a distinct developmental profile relying more heavily on orthographic to semantic mappings

and learning these mappings more rapidly even than orthographic to phonological mappings. These models provide support for the position that reading, regardless of orthographic structure, can be supported by the same computational system operating over distributed representations of phonological, semantic and orthographic information, with orthographic systems affecting how those representations interact within the reading system.

Cognitive neuroscience studies have shown that, irrespective of the orthographic system, the neural architecture supporting reading across populations of different languages spans many of the same key brain regions (Tan, et al., 2005; Bolger et al., 2005; Das et al., 2011; Nakamura et al., 2012), that is, the left lateralised brain networks shown to support spoken language processing (Devauchelle et al., 2008) that is largely in place from two months of age (Dehaene-Lambertz et al., 2010). There is, however, also neuroscientific evidence to support the existence of two distinct paths in the reading system (e.g. Jobard et al., 2003; Price, 2010; Richardson et al., 2011). Largely derived from studies conducted on alphabetic literate participants, neuroimaging data demonstrates that reading recruits both a dorsal (orthography to phonology) path and a ventral (orthography to semantics) path, though models of this dual stream have been implemented for alternative orthographies, such as Ueno & Lambon Ralph's (2013) Japanese model.

Studies that have compared neural activation in literate populations that differ in the transparency of the orthography on which they were trained have revealed differences in activation of regions associated with orthography to phonology and orthography to semantics pathways. A study by Paulesu et al. (2000) used Positron Emission Tomography (PET) to compare brain activity of English (deep alphabetic) and Italian (shallow alphabetic) literates when reading. They observed greater activation of regions associated with the ventral pathway in English literates, a finding that fits with suggestions that deeper orthographies rely more heavily on direct orthography to semantic mappings. Kiyosawa et al. (1995) conducted a within participant analysis comparing cerebral blood flow when Japanese literates read stimuli presented in either Kanji (logographic) or Kana (syllabic) script. Their analysis showed a bias towards ventral processing when reading in Kanji and a dorsal bias when reading Kana, supporting evidence for changes in orthographic transparency altering dependence on a direct orthographic to semantic route.

Neuroimaging studies have also investigated changes to processing over the course of literacy training, although this research is again dominated by studies of alphabetic literates. Current

data indicate a progression over the course of training in English literates from an initial bias towards use of the dorsal (orthographic to phonological) path to later dominance of the ventral (orthographic to semantic) path in proficient readers (Pugh et al., 2001; Shaywitz et al., 2002). These data have motivated contrasting interpretations. For example, Pugh et al. (2000; 2001) suggested that in proficient readers the dorsal path is only recruited for slow analytic mapping from orthography to phonology, while Levy (2008; 2009) claimed that the dorsal path is primarily recruited for non-word reading, as in the DRC model (Coltheart et al., 2001). Such a perspective however seems unlikely given recent data showing that the dorsal path is involved at both early and late stages of written word processing (Richardson et al., 2011).

In summary, neuroimaging and computational modelling data suggest that reading, irrespective of the orthographic system, is supported by a similar underlying neural architecture that incorporates pathways between brain regions involved in processing orthographic, phonological and semantic information. These studies also suggest that orthographic transparency interacts with exposure in impacting the nature of processing in such networks, particularly the distribution of labour across processing pathways.

Effects of literacy on representational structure

There is substantial evidence that learning an orthographic system affects processing of spoken forms of words, thus developing a mapping from orthography onto phonology affects the structure of those phonological representations (Huettig & Mishra, 2014; Morais & Kolinsky, 2001; Petersson, Ingvar & Reis, 2009). The type of orthographic system may exert a substantially different effect on this phonological restructuring, though it is once again difficult to isolate behavioural effects of orthographies from linguistic or socio-cultural factors. Numerous studies have demonstrated that learning to read an alphabetic language leads to increased performance on tasks that test participants' ability to detect and manipulate individual phonemes in speech (referred to as phonological awareness tasks). These effects have been demonstrated in both child (Alcock et al., 2010; De Jong & Van der Leij, 1999; Hulme, Snowling & Caravolas, & Carroll, 2005; Treiman & Zukowski, 1991) and adult (Adrian, Alegria & Morais, 1995; Loureiro, Willadino Braga, Souze, Queiroz & Dellatolas, 2004; Morais, Cary, Alegria & Bertelson, 1979; Scliar-Cabral, Morais, Nepomuceno, & Kolinsky, 1997) populations. There is also growing evidence that literacy affects changes to phonological processing during online speech processing (Hoonhort et al., 2011; Huettig &

Mishra, 2014; Huettig, Singh, & Mishra, 2011; Reis & Castro-Caldas, 1997; Serniclaes et al., 2005), and a computational model implementing mapping from speech to visual and semantic information (Smith, Monaghan, & Huettig, 2014a) simulated these changes in processing as increasingly fine-grained phonological processing during online speech processing as a consequence of advanced literacy in alphabetical orthographies.

Three theoretical models have been proposed for how learning orthographic mappings may affect phonological processing. First, it could be that orthographic representations are activated online during speech processing, and interconnections to and from orthography then influence phonological processing (online activation hypothesis) (e.g. Ziegler & Ferrand, 1998), consistent with Price and Devlin's (2011) interactive account of ventral processing in reading. Second, learning to map between orthographic and phonological representations may lead to a restructuring of phonological processing (e.g. Muneaux & Ziegler, 2004; Ziegler & Goswami, 2005). Third, it could be that literacy results in both online activation of orthography *and* phonological restructuring (e.g. Dehaene et al., 2010). Dehaene et al. (2010) used fMRI to compare brain activity in illiterate and alphabetic literate populations when processing spoken words. In literate, but not illiterate, populations they observed activation of brain regions associated with orthographic processing (left occipital temporal cortex), as well as a near doubling of activation in the planum temporale, a region associated with phonological processing (see also Monzalvo & Dehaene-Lambertz, 2013). This support the hypothesis that orthographic representations are activated online and that there are changes to phonological representations as a consequence of literacy training, though the temporal resolution of fMRI may not be reflecting only online language processing of the few hundred milliseconds of the spoken word but also potentially longer-term side-effects not critically involved in early stages of phonological encoding.

Further evidence in support of the phonological restructuring hypothesis can be found in a recent ERP study by Perre et al., (2009). The study demonstrated that orthographic effects observed during lexical decision tasks, approximately 330ms post word onset, could be localised to neural populations classically associated with phonological processing (left BA40). They found no evidence for effects being generated by brain regions associated with orthographic processing. Perre et al., interpret these findings as strong evidence in support of changes to phonological processing during early stages of online speech processing as a consequence of literacy training being due to a restructuring of phonological processing regions rather than online activation of orthographic representations. This conclusion is

further supported by a Pattamadilok et al. (2010) who used TMS to modulate activity either in regions associated with phonological processing (left supremarginal gyrus) or regions associated with orthographic processing (left ventral occipitotemporal cortex) as participants completed a lexical decision task. They observed that orthographic consistency effects were eliminated when phonological regions were manipulated, yet no effect was observed when TMS was applied to orthographic processing regions.

Current cognitive neuroscientific data suggests that exposure to alphabetic literacy training affects the structure of phonological processing regions involved in online speech processing. This raises the question as to whether development of literacy in other orthographic systems exerts the same effects on phonological processing. Psycholinguistic grain size theory (Ziegler & Goswami, 2005) predicts that the nature of correspondence between graphemes and phonological units for a given orthographic system will define the impact of literacy training on phonological processing. On this basis alphabetic literates will exhibit finer grain effects on phonological processing than logographic literates due to the regular systematic relations between individual graphemes and phonemes in alphabetic systems, hence the phonological restructuring proposed to result in phoneme awareness abilities and finer-grained processing of speech in visual world processing tasks is likely to be observed primarily as a consequence of literacy for alphabetic orthographies.

Computational modelling work has provided an explicit description of how the granularity of processing can be influenced by literacy training for an alphabetic orthography. Harm & Seidenberg (1999) trained a phonological attractor network to sustain phonological representations over time for English monosyllabic words (illiterate model). They also created a "literate model" that was in addition trained to map from orthographic representations to the same set of phonological representations. The literate model was more robust in restoring corrupted phonemes and phonological features within phonemes, indicating that phoneme-level representations were stronger as a consequence of literacy. This advantage was due to the literate model developing stronger connection weights than the illiterate model between processing units that were representing the same phoneme segment, as a consequence of learning the systematic relation between phonological and orthographic representations. This approach was extended to explore whether the same effects are observed irrespective of the orthographic structure on which the networks are trained. Smith, Monaghan & Huettig (2014b) trained literate models identical to those used in Harm & Seidenberg, (1999) on a transparent orthography representing English, or an opaque

orthography, where orthographic and phonological representations of words were the same as for the transparent model, but the pairings between the representations was randomised. They observed that the structure of the orthographic system differentially impacted on phonological processing, with networks trained on a transparent orthography displaying finer grained processing than those trained on an opaque orthography.

It is becoming increasingly difficult to test literate populations that are not exposed to alphabetic systems due to the increasingly widespread use of English, and the evolution of simultaneously used alphabetical or alphasyllabic writing systems alongside more opaque orthographies, e.g., the increasing use of Pinyin accompanying traditional writing for Mandarin Chinese (Cheung & Ng, 2003). There is, however, some evidence to support the predictions of psycholinguistic grain size theory and the predictions of the computational work described above. An early study by Read et al., (1986) tested adult literates either literate only in logographic Chinese script or literate in both alphabetic and logographic scripts. Their results showed that only the group literate in both the alphabetic and logographic scripts were able to add or delete individual consonants in spoken Chinese words. Similarly, De Gelder & Vroomen (1992) tested a group of Dutch literates, a group of Chinese literates and a group literate in both Dutch and Chinese on their ability to distinguish between /ba/ and /da/ drawn from a 9 step continuum. They showed that alphabetic literates (Dutch) and literates of two scripts (Chinese [logographic] and Dutch [alphabetic]) displayed sharper phonological boundary precision than logographic (Chinese) literates. Further, Cheung & Chen (2004), testing participants' ability to perform sound matching and primed shadowing tasks, and Shu et al. (2008), who tested phoneme onset awareness, demonstrated that performance on these phonological awareness tasks in Chinese literates coincided with exposure to Pinyin tuition. However, Kidd et al. (2014) showed that logographic literates can show sensitivity to phonological structure in speech gating and non-word repetition tasks (Kidd et al., 2014), though it was unclear the extent to which these participants were exposed to Pinyin, however, the effect of orthography appears to be one of degree rather than a qualitative change in processing. Relatedly, Brennan et al., (2013) used fMRI to compare Chinese and English literate adults and childrens' brain activity when performing rhyme judgement tasks. They observed differences over the course of development in phonological processing regions (superior temporal gyrus) only in comparisons between English speaking adults and children.

Together the literature offers substantial theoretical and empirical support for an effect of orthographic transparency on phonological processing.

Modelling the effects of orthographic systems

The goals of this study are as follows. Firstly, we determined the range of the triangle model of reading as a universal architecture able to implement reading across each of the world's major orthographic systems. Secondly, after establishing its adequacy to simulate reading across orthographic systems, we investigated the effect of different orthographic systems on the way in which the reading system develops as a consequence of literacy. We hypothesised that the varying systematicity between orthography and phonology and orthography and semantics would result in distinct patterns of division of labour along pathways between orthographic, phonological, and semantic representations, and that these are likely to vary during the course of reading development (Harm & Seidenberg, 2004). Furthermore, we hypothesised that the model trained on an alphabetic orthography would demonstrate finer-grained phonological processing than the model trained on logographic systems. Alphasyllabic orthographies were predicted to lie somewhere in between. In addition, given that orthographic effects have been observed on phonological processing which are driven by the extent to which systematic relations are embedded in the orthography to phonological representations, it seems likely that analogous effects of literacy on *semantic* processing could be observed for orthographic systems in which systematic relations exist between orthography and semantics, such as are represented in the semantic radicals of traditional Chinese characters. Finally, words that are unrelated in phonological or semantic dimensions may share relationships in their orthographic form, we examine whether in an interactive model of reading whether such relationships affect the structure of representations beyond the orthographic domain. To our knowledge, there are as yet no such studies that have investigated either of these issues.

To examine each of these issues we trained a connectionist neural network model based on Harm & Seidenberg's (2004) implementation of the triangle model of reading, but using artificial corpora consisting of orthographic, phonological and semantic representations. Constructing artificial corpora ensures that relations both within and between representational domains are controlled (see Hirshorn & Fiez (2014) and Plaut et al. (1996) for similar justification). Consequently, controlling semantic and phonological representations and mappings across orthographic systems ensures any observed differences in behaviour or

processing are driven by differences in the orthographic structure. In behavioural studies, understanding the effects of orthographic transparency is compromised by difficulties in isolating its effects from confounding linguistic, socio-economic or socio-cultural factors (such as differences in teaching methods, extent of language exposure in different modalities, or other social economic or cultural factors that influence differences in literacy exposure). Using a computational framework can isolate the information processing constraints that apply from different orthographies, which have an influence on the way in which mappings can be formed during literacy development.

There are several modelling paradigms that have effectively simulated a wide range of behavioural phenomena associated with reading (Coltheart et al, 2001; Harm & Seidenberg, 2004). In terms of precision of match between behaviour and model performance, the most advanced of these are a hybrid series of models, termed CDP+ and CDP++. These models combine the dual-route framework in terms of incorporating both a mapping from written to spoken forms of words via a set of lexical representations (where each word is represented by a single processing unit) along with a connectionist implementation of an orthographic to phonological pathway (Perry, Ziegler, & Zorzi, 2007, 2010). This approach has been successfully extended to cover alphabetic orthographies other than English (Perry, Ziegler, & Zorzi, 2014a, 2014b). However, we selected a connectionist model as the framework for our investigations for several reasons. Principally, we wanted to explore the way in which a model learned to solve the task, given the constraints of the mappings that the model was required to form, and the extent to which this involved division of processing across different pathways mapping between orthography, phonology and semantics. An alternative would be to implement a trained model representing the lexicon (including semantic representations) and train only the orthography to phonology direct pathway, as in the CDP+/CDP++ tradition. However, then we would be introducing assumptions into the relative role of these pathways, thereby reducing the explanatory power of the model. Furthermore, the trajectory of learning of the model was a critical issue. Though in the CDP models, learning is instantiated in the orthography to phonology mappings, it is not yet clear how this interacts with learning of the lexical pathway and how that might be affected differentially by alternative orthographies. We contend that our approach is not incompatible with these other modelling traditions in reading research, but that fewer assumptions of architecture and fewer constraints on processing enable a valuable first step in understanding how different

orthographies involve and modulate processing between different representational forms for words.

The orthographies selected for modelling represented each of seven orthographic structures: shallow alphabetic; deep alphabetic; alphasyllabic; consonantal; syllabic; logographic semantically opaque; logographic semantically transparent. Critically, identical semantic and phonological representations and mappings were used across all orthographic systems, and all training and testing parameters of the model were controlled to ensure that observed behavioural effects of the model stemmed only from effects of the orthographic representations, and the extent to which they reflected aspects of phonological and semantic structure. This close control and comparison of models of multiple orthographies enable a test of theoretical and neuroscientific accounts of division of learning rates of reading fluency and reading comprehension affected by orthography, and the extent to which the architecture of the reading system in terms of direct and indirect routes and division of labour can explain apparent behavioural distinctions in reading performance.

2. Multiple orthographies in the triangle model of reading

Architecture

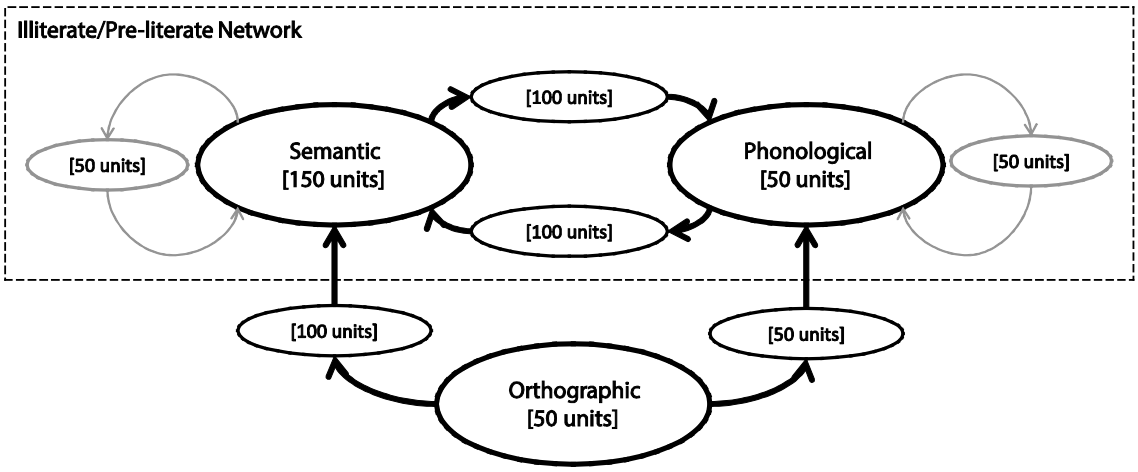

Figure 1: Model Architecture

Model architecture (Fig. 1) was closely matched to Harm and Seidenberg's (2004) implementation of the triangle model of reading, differing only in terms of the number of units within each of the layers (see appendix). The illiterate/pre-literate network consisted of

an interconnected phonological and semantic layer, to simulate spoken language processing. The semantic layer comprised 150 units. These units were fully connected to a set of 25 semantic clean-up units, which were fully connected back to the semantic layer. These clean-up units were included to ensure that stable semantic representations could be formed in the semantic layer. The semantic layer was fully connected to a hidden layer of 100 units which was in turn fully connected to a phonological layer consisting of 50 units. The phonological layer was also connected to a set of 25 phonological clean-up units which were connected back to the phonological layer. The phonological layer was also connected to the semantic layer via another hidden layer consisting of 100 units. The reading model built on the pre-literate/illiterate network with the addition of an interconnected orthographic layer. The orthographic layer consisted of 50 units, which was connected to the phonological and semantic layers via two hidden layers of 50 and 100 units, respectively. Numbers of units in the hidden layers were determined by pilot studies to determine the minimum number of units required to form mappings between representations. One exception being that in all pilot simulations the number of hidden units in the orthography to phonology path was half that of the number in the hidden layer connecting orthography to semantics, this reflects the architectural assumptions implemented in Harm & Seidenberg (2004). A bias unit was fully connected to each layer within the network.

Representations of words

Artificial corpora of words were constructed consisting of 500 unique items. Each item represented a unique monosyllabic word and was assigned a phonological, semantic, and orthographic form. For each item, the semantic representation was unique, but there was some overlap between the phonological and orthographic representations, to simulate homophones in the corpus. Homophones were included to ensure a distinction between syllabic and logographic systems at the monosyllabic level. Eight different artificial corpora were generated for each of the seven orthographic systems to be simulated.

Semantic Representations

Semantic representations were encoded by a 150 unit binary feature vector, with p(active feature) = 0.1. Prototypes similar to those used in Dilkina, McClelland & Plaut, (2010) were used to construct an artificial semantic taxonomy; consisting of 2 high-level semantic categories, each with 5 sub-categories with 50 items per sub category (see Figure 2). Items were both more likely to share semantic properties with items within the same sub-category

[p(within sub-category) = 0.267] than items within the same high level category [p(within high category) = 0.133], and were more likely to share properties with the same high level category than items outside their high level category [p(outside high category) = 0.040].

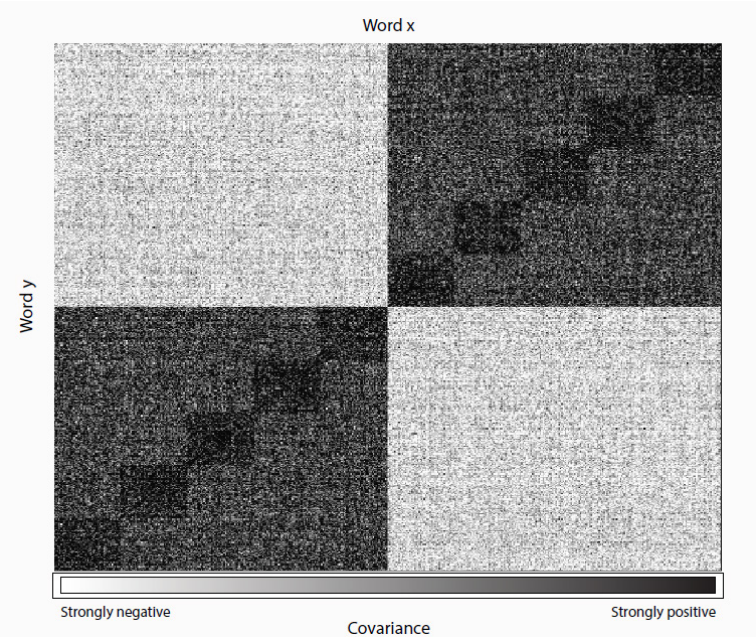

Figure 2: Matrix displaying covariances among semantic representations in the training corpus. Dark colours indicate strong positive covariance, light colours indicate strong negative covariance.

Phonological Representations

Phonological representations were 5 phonemes in length and of the form CCVCC[1]. The phonological layer consists of 5 phoneme slots, organised CCVCC, with each slot 10 units in length (total no. units = 50). A phoneme inventory consisting of 5 vowels and 10 consonants was constructed. Each phoneme was encoded by a unique 10 unit phonetic feature vector with p(active) = 0.5. Phonemes were pseudo randomly sampled from the phoneme inventory to construct words while ensuring across all words each phoneme was used an equal number of times in each phoneme slot (ignoring homophones). Four hundred and fifty distinct phonological representations were constructed, and then a further 25, distinct from the 450,

[1] Languages vary in terms of whether or not they contain consonant clusters, and also regarding word length in terms of how representative a monosyllabic subset of the language would be. In this respect, we are abstracting away from the properties of particular languages, but incorporating sufficient complexity within the phonological forms to permit overlap and distinctiveness in the representations, reminiscent of the patterns found across a wide set of languages. We did not implement different phonological characteristics across orthographies because we wanted to keep the semantic and phonological representations the same, in order to isolate the contribution of variations in the mappings between representations to reading performance and language processing.

were included twice, mapping to different orthographic and semantic representations, to simulate presence of 25 homophones in the language.

The phonological representations of 50 non-words were also constructed by randomly sampling phonemes from the phoneme inventory. Non-words were not used in training but used to test the model post-training to examine alphabetic, alphasyllabic and consonantal networks' ability to generalise to novel forms (see section Non-word Reading).

Orthographic Representations

Seven forms of orthographic structure were implemented, based on the descriptions of different writing systems provided in Comrie (2013). All orthographic representations consisted of a 50 unit binary feature vector where each feature had p(active) = 0.5.

Alphabetic: In alphabetic systems the basic unit of representation is the phoneme. The orthographic layer was defined in terms of 5 letter slots, organised in a CCVCC structure to match that of the phonology, where each slot contained 10 units. All words consisted of 5 letters taken from an alphabet of 15 letters, where 5 letters represented vowels and 10 represented consonants.

Alphabetic Shallow: In the shallow alphabetic system there was perfect correspondence between occurrence of a letter and a phoneme. Each letter within the alphabet was assigned a phoneme. Orthographic representations were constructed using this regular, transparent mapping. Reflecting controls on phonological representations the occurrence of each letter was controlled for letters representing consonants and letters representing vowels. 25 homographs were included in the alphabetic shallow corpus reflecting the presence of 25 homophones in the phonology. Orthographic representations for each of the 50 non-words in the phonology were constructed using corresponding phoneme to letter mappings.

Alphabetic Deep: In the deep alphabetic system, orthographic representations were constructed using the same procedure as outlined for the shallow alphabetic system with one variation, where 20% of mappings for each vowel were irregular. This was implemented by replacing within the alphabetic shallow representations on 20% of occasions for each vowel the letter assigned to the given vowel with one of the other 4 letters representing vowels. Pseudo-randomisation ensured that irregular mappings occurred an equal number of times for each vowel, with each alternative vowel used as the replacement an equal number of times this ensured the occurrence of letters representing each vowel was controlled. Also, reflecting

controls on consonants in the phonology each letter representing each consonant occurred an equal number of times within the corpus. Controls ensured that irregular mappings did not increase the number of homographs embedded within alphabetic deep corpora compared to alphabetic shallow systems. In total 25 homographs were embedded in each alphabetic deep corpus. Orthographic representations of non-words were constructed using the regular phoneme to letter mappings defined for each alphabetic deep corpus.

Consonantal: Within a strictly consonantal system only consonants are represented (although in many consonantal scripts it is possible to add diacritics to indicate vowels). Within our implementation of the consonantal system the orthographic layer had 4 letter slots, organised as CCCC. An alphabet of 10 letters was constructed, each representing one of the 10 consonants in the phoneme inventory, and each letter was represented in terms of 12 features. Two of the 50 units in the orthographic input were therefore always inactive for the consonantal writing system. The use of 12 letter features for each letter slot, instead of 10, was used to ensure that the overall complexity of the orthographic input was similar to that of the alphabetic system simulations. As for the alphabetic simulations, the number of occurrences of each letter in each position was balanced. As vowels are not represented in this system the number of homographs is larger than the number of homophones. Consonantal corpora contained on average 34 homographs ($\mu = 33.6$, $\sigma = 2.91$). Orthographic representations were also created for consonantal systems using the consonant to letter mappings as defined above.

Alphasyllabic: Within alphasyllabic systems a basic grapheme indicates a consonant, with a diacritic added to indicate its combination with a particular vowel (in contrast to consonantal languages vowels must be indicated). In our implementation of the alphasyllabic system the orthographic layer consisted of 4 slots, organised C[CV]CC, with C slots defined by 12 units and the [CV] slot defined by 14 units, to simulate a diacritic added to a consonantal character. Prototypes were produced for each consonant and used to create 5 unique 14 unit vectors for each consonant, each of which represented the consonant in combination with one of the 5 possible vowels. Prototypes ensured that 12 features were shared between representations of the same consonant with $p = 0.8$, while the remaining two features were shared with $p = 0.5$. This ensured that each representation of a given consonant was more similar to other representations of the same consonant and the signal indicating the vowel was distributed across the entire consonant-vowel slot. Due to the regularity of the mappings within this system the 25 homophones embedded in the orthography result in 25 homographs within

each alphasyllabic corpus. Orthographic representations were also created for all 50 non-words using the phoneme to grapheme mappings specific to each alphasyllabic corpus.

Syllabic: In syllabic systems a distinct grapheme encodes each syllable. No sub-component of the orthography can be identified as denoting distinct sub-syllabic segments. In our implementation of a syllabic system the orthographic layer consisted of single slot defined by 50 units. A unique 50 unit binary feature vector [p(active) = 0.5] was created to form the orthographic representation of each syllable (CCVCC) within the corpus. There were therefore 25 pairs of homographs in this writing system, reflecting the 25 homophone pairs in the phonology.

Logographic Semantically Transparent: In many logographic systems components of the orthography provide probabilistic information regarding the semantic category of a word. To explore the effects of this property we constructed a logographic system in which there was greater overlap of orthographic features between words within the same semantic sub-category than words within the same semantic higher level category, and greater overlap of orthographic features between words within the same higher level category than between words in different high level semantic categories. Thus, aspects of the orthography approximated the semantic structure of the items. Logographic semantically transparent orthographic representations consisted of a unique 50 unit binary feature vector [p(active) = 0.5]. The first 30 features were more likely to be shared with items within the same semantic category. These 30 features were split into two sets of 15 features, each subdivided in into 5 sets of 3 features, with each subset assigned a given semantic sub-category. The probability that a given orthographic feature was active was dependent on the words distance from the semantic category it was attributed to: p(active) = 0.8, for a feature assigned to the same semantic sub-category as the given word; p(active) = 0.6 for a feature assigned to the same high-level semantic category as the given word; and p(active) = 0.36 for a feature assigned to a high-level semantic category that differed from that of the given word.

Logographic Semantically Opaque: In order to isolate the effects of semantic transparency from logographic structure we also constructed logographic semantically opaque systems. In the case of monosyllabic words semantically opaque logographic systems would differ from syllabic systems only in that two items with identical phonological representations that differ in their meanings will have distinct orthographic representations in the logographic language, but would be identical in phonology and orthography in the syllabic system. Within our

implementation of logographic and syllabic systems the orthographic layer consisted of a single 50 unit slot. For logographic semantically opaque systems a unique 50 unit binary feature [p(active) = 0.5] vector encoded each word within the corpus, thus, there were no homographs in this writing system.

Training

The model was trained to map between phonological, semantic, and orthographic representations for all words in the training set. Each training trial ran for 4 units of time, where in each unit of time activation passed fully from one layer to the following layer. These time units were each divided into 4 samples (time steps), with an integration constant of 0.33 which meant that activation passed from one layer to another gradually and accumulatively, with a full pass of activation every 4 time steps (see appendix).

The training regimes were identical across each orthographic system. There were two stages to the training. During pre-literacy training, the model was exposed to phonological and semantic representations of words, to simulate the child's exposure to spoken language prior to learning to read. Literacy training comprised learning to map orthographic forms of words onto phonological and semantic representations, whilst also maintaining performance on the pre-literate tasks. This was to simulate the interleaving of exposure to reading tasks and spoken language whilst children are learning to read.

Pre-literacy training consisted of four tasks which varied in their probability of occurrence: phonological retention task, p = 0.1; semantic retention task, p = 0.1; speech comprehension, p = 0.4; and speech production, p = 0.4.

For phonological retention trials, the model was given the word's phonological representation, and was trained to stably maintain that representation over the 12 time steps. Words were randomly selected from the corpus. Its phonological representations was clamped to the phonological layer for time steps 0 – 7. The target (phonological representation of the selected word) was provided from time step 8-12 and error back propagated. For this, and the following tasks, the presentation of the target was only made at the point when activation had passed fully from the layer provided with the input to the layer representing the required output (i.e., 8 time steps).

Semantic retention trials required the model to sustain semantic representations in the semantic layer over time. A randomly selected semantic representation from the training

corpus was clamped to the semantic layer for time steps 0 – 7. The target was then provided from time steps 8-12 and error back propagated.

Speech production trials trained the network to map from semantics to phonology, in order to simulate spoken word production. A randomly selected word was taken from the corpus for each trial. Its semantic representation was clamped to the semantic layer for time steps 0 – 7. For time steps 8 – 12 the word's phonological representation was provided as target to the phonological layer and error back propagated.

Speech comprehension trials trained the network to map from phonology to semantics. A randomly selected word's phonological representation was clamped to the phonological layer for time steps 0 – 7. For time steps 8 – 12 the word's semantic representation was presented to the semantic layer as a target and error back propagated.

Note that the pre-literacy training was identical for all simulations of the distinct orthographies, because the phonological and semantic representations were identical, just the orthographic forms varied. Networks were trained on a total of 150,000 pre-literacy trials before the onset of literacy training. This ensured that networks were able to perform all tasks to a high degree of accuracy before the onset of literacy training: for the phonological retention and semantic retention tasks, accuracy was 100%; for the speech comprehension task, accuracy was 95%, (this was the expected maximum because 5% of the training patterns were indistinguishable because of the inclusion of homophones in the training set); and for speech production, accuracy was 100%. In each case, accuracy was determined as the percentage of items within the training corpus for which output was closest to the target.

Literacy training consisted of the four pre-training tasks [phonological retention task, $p = 0.05$; semantic retention task, $p = 0.05$; speech comprehension, $p = 0.25$; speech production, $p = 0.25$] and an additional reading task ($p = 0.4$).

Reading trials trained the network to map from orthography to simultaneously produced phonological and semantic representations. For time steps 0 – 7 the orthographic representation of a randomly selected word was clamped to the orthographic layer. During time steps 8 – 12, the phonological representation of the word and the semantic representation of the word were presented to the phonological layer and semantic layer of the network respectively, as a target and error back propagated. Networks were trained on a further 100,000 training trials with pre-training and reading trials randomly interleaved. At

the end of literacy training all simulations performed pre-training tasks to the same level of accuracy as displayed prior to literacy training. Performance on reading tasks is detailed in later sections of this paper. Words were randomly selected from the training corpus with equal probability for use on training trials. Continuous recurrent backpropagation (Pearlmutter, 1989) was used to train the model with a learning rate of 0.1 (see appendix). Connection weights within the network were initiated with random weights in a uniform distribution in the range [-1, 1]. Eight simulation runs, each initiated with a different random seed, were trained for each orthographic system. All training and testing parameters were controlled in the same way across orthographic systems.

3. Examining the Effects of Orthographic Transparency

The preliterate training environment was controlled across simulations of each orthographic system, because phonological and semantic representations of words were identical across each orthography, thus there was no variation according to orthographic system in the pre-literacy training results. Furthermore, all simulations attained maximal levels of performance on all pre-literacy training tasks prior to the onset of literacy training.

We first report performance on the model's acquisition of reading from the different orthographic systems. A key issue in computational modelling of (alphabetic orthographies) is the ability of the model to generalise to previously unseen items, such as in a non-word pronunciation task. Original criticism of connectionist models of reading focused on poor generalisation performance (e.g., Seidenberg & McClelland, 1989; Coltheart et al., 2001), though subsequent connectionist models have effected this generalisation (e.g., Plaut et al., 1996; Harm & Seidenberg, 1999). Nevertheless, nonword reading is a benchmark effect for any simulation of trained reading. We therefore test the model's ability to generalise pronunciation to previously unseen words. Then, we investigate the division of labor that emerges across the two pathways (orthography to phonology and orthography to semantics) in the model for reading pronunciation and comprehension. Finally, we provide an analysis of the differential effects of the orthographic systems on changes to the way in which the model represents structure in the phonological and semantic systems. These analyses give rise to hypotheses about the varied effects of literacy on phonological and semantic processing across writing systems.

Reading Acquisition

For the reading acquisition tasks, we report performance on reading aloud pronunciation, where the model's performance on phonological output given orthographic input is assessed. Then, we describe the model's performance on the reading comprehension task, where for given orthographic input, the model's semantic output is appraised. These results will demonstrate the extent to which orthographic systems affect the trajectory of reading development, and the relative speed of learning to read for pronunciation and for comprehension.

Acquiring phonological decoding abilities under different orthographies

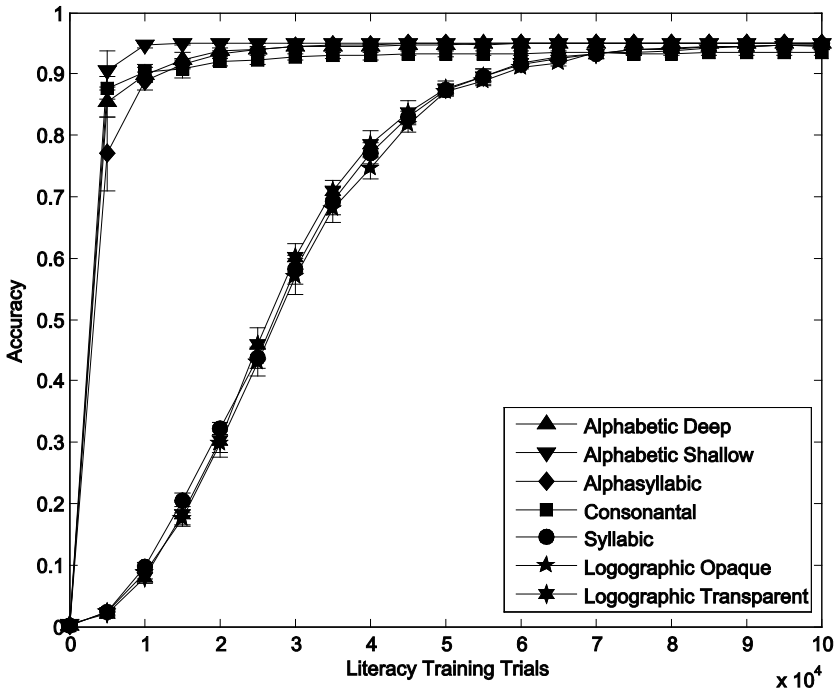

Figure 3: Phonological decoding accuracy during literacy training for each orthographic system.

Each simulation was tested on its ability to accurately produce the phonological representation of a word when presented with its orthographic representation during training, at 5000 training patterns intervals. Performance was recorded at the 12th time step in the phonological layer of the network. The cosine distance between the output activation and all phonological representations in the training corpus was calculated. A word was recorded as accurately produced if its phonological representation was closest to the activation recorded

in the phonological layer. Figure 3 displays the accuracy for all words in the set of 8 simulation runs for each orthographic system during the first 100,000 literacy training trials. A clear distinction emerges from the onset of literacy training between sub-syllabically transparent systems and the remaining systems in decoding acquisition rate with sub-syllabically transparent systems rapidly reaching accuracy levels exceeding 70% after 5000 trials, whereas logographic and syllabic systems require approximately 40000 trials to reach similar levels of performance.

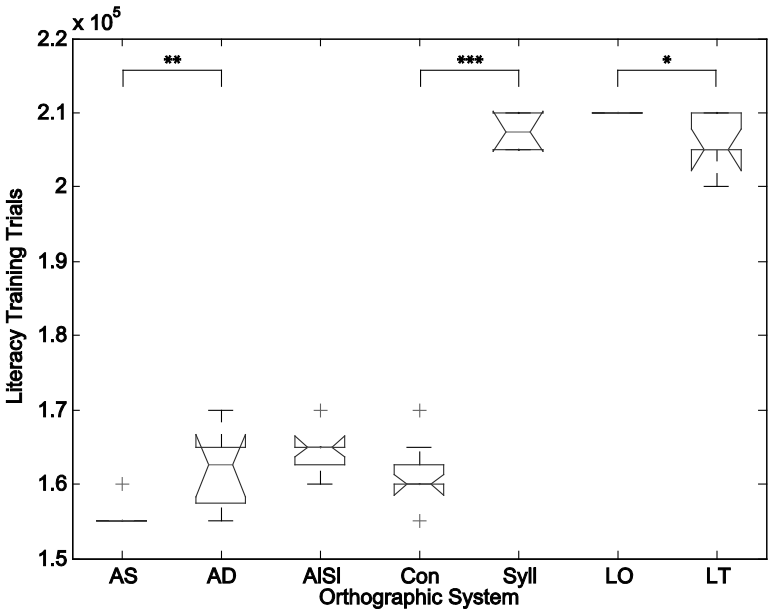

Figure 4: Literacy training trials required to exceed 90% accuracy on phonological decoding task. [AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent] [2].

Figure 4 records the number of literacy training trials required for each type of orthography simulation to reach 90% accuracy on orthographic to phonological mappings. Due to the number of homographs within consonantal corpora certain consonantal networks were only capable of achieving 92% accuracy in reading comprehension. A threshold of 90% accuracy

[2] Box plots were constructed using MATLAB (Version 7.9.0.529) function boxplot. On each box within the box plot, the central mark indicates the median (q2) with the upper and lower edges indicating the 25th (q1) and 75th (q3) percentiles. Extreme data points that are not considered outliers [i.e. outlier > q3 + 1.5(q3 – q1) or outlier < q3 - 1.5(q3 – q1)] are indicated by the whiskers that extend from each box. Outliers are plotted individually. Medians differ significantly if their intervals do not overlap. Interval endpoints (i.e. q2 ± 1.57(q3 – q1)/√n) are indicated by the notches on the side of each box.

was therefore chosen to ensure comparisons across systems were conducted at a level of proficiency attainable for all networks.

A one-way ANOVA confirmed that networks differed in their reading for production acquisition rates, $F(6,49) = 445.5$, $\eta^2 = 0.982$, $p < .001$. Six two-sample t-tests were performed (Alphabetic Shallow, Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic Transparent, Logographic Opaque) and corrected for multiple comparisons using a Bonferroni correction, to examine how individual orthographic systems differed from one another in their reading for production acquisition rates. This analysis revealed that shallow alphabetic networks reached 90% accuracy before alphabetic deep networks ($M = -6250$, $SD = 3952.8$, $t(14) = -3.162$, $p = 0.007$). Alphabetic deep networks did not differ from alphasyllabic ($M = -2500$, $SD = 4381.3$, $t(14) = -1.141$, $p = 0.272$), while alphasyllabic did not differ from consonantal networks ($M = -2500$, $SD = 3867.2$, $t(14) = 1.616$, $p = 0.128$). However, consonantal networks achieved 90% accuracy in orthographic to phonological mappings faster than syllabic networks ($M = -2500$, $SD = 3659.6$, $t(14) = -25.28$, $p < 0.001$). While there was no difference in acquisition rates displayed between syllabic and logographic transparent ($M = 1250$, $SD = 3133.9$, $t(14) = 0.797$, $p = 0.438$). There was however a marginal difference between logographic transparent and logographic opaque networks ($M = -3750$, $SD = 2500$, $t(14) = -3$, $p = 0.010$), although this was not significant when correcting form multiple comparisons.

The model replicates known findings that orthographic transparency increases the rate of phonological decoding acquisition. The increased acquisition rate of alphabetic shallow networks over alphabetic deep networks supports such conclusions from studies that control for exposure to training while witnessing reduced performance on word and non-word reading tasks in populations learning deeper alphabetic systems (Finnish vs English: Goswami, Gombert & De Barrera, 1998; Seymour, Aro & Erskine, 2003; French vs English: Bruck, Genesee & Caravolas, 1997; Welsh vs English: Hanley, Masterson, Spencer & Evans, 2004). Although comparing the number of trials required in order to achieve 90% accuracy on decoding tasks did not distinguish between alphasyllabic and alphabetic deep trials, there is a suggestion of a difference as indicated by the figure (see figure 4) that alphasyllabic networks were delayed in comparison to alphabetic deep networks at earlier stages of training. This aligns with the empirical data reported in Nag (2007) that describes delayed decoding acquisition in populations learning Kannada (alphasyllabic) in comparison to

populations learning English (deep alphabetic). Also, as expected given the empirical data, logographic systems displayed the slowest decoding acquisition rates. In contrast however to data reported in Asfaha et al. (2009) that showed increased acquisition rates in populations learning Ge'ez scripts (alphasyllabic) compared to those learning Latin scripts (alphabetic), in our simulations syllabic systems displayed no decoding advantage over any other system. The model's failure to capture the observations reported in this study are likely due to the limitations of modelling only monosyllabic words and assumptions regarding the structure of phonological input to the system, we return to this issue in the General Discussion. Logographic transparent networks were marginally distinguishable from logographic opaque networks in their ability to learn orthographic to phonological mappings, although as authors we are unaware of any existing empirical data examining the effects of semantic transparency on phonological decoding in logographic systems, the current model suggests that systematic relations between orthography and semantics may increase the rate of acquisition of decoding abilities. We presume that this is due to systematic relationships between orthography and semantics increasing the rate at which orthographic to semantic mappings are learnt in logographic transparent networks, thus decoding acquisition can benefit from early access to information from the orthographic to phonology via semantics path. We will return to this issue in our examination of the division of labour within networks later in this chapter.

Acquiring reading comprehension abilities under different orthographies

The trajectory of learning orthography to semantic mappings was also examined (reading comprehension). To test accuracy on these mappings, the model's output in the semantic layer was analysed in the final time step (ts = 12) for a reading trial, and was compared to all semantic representations within the corpus. If activation in the semantic layer was closest to the target word then the item was recorded as read accurately. Figure 5 displays the accuracy of networks on this reading for comprehension task, as a proportion of items read within the corpus, over the course of literacy training. The model successfully learns to accurately map from orthographic to semantic representations for all words in the training corpus (allowing for variation in the number of homographs across systems) for all orthographic systems. In this respect it is comparable to previous models of reading and semantics (Plaut et al., 1996; Harm & Seidenberg, 2004; Monaghan et al., 2004).

Figure 5 shows, as was apparent in rates of phonological decoding acquisition, that a distinction can be made between reading comprehension acquisition rates displayed by

systems that possess sub-syllabic phonological structure in their orthography and those that do not, with sub-syllabic phonological transparency leading to faster rates of acquisition in reading comprehension. The system with greatest phonological transparency (alphabetic shallow) appears to demonstrate the fastest rate of acquisition. As can be observed from figure 5 phonologically transparent networks do not reach 100% accuracy in reading comprehension, this is due to the presence of homographs for which the semantic target can't be determined from the orthography.

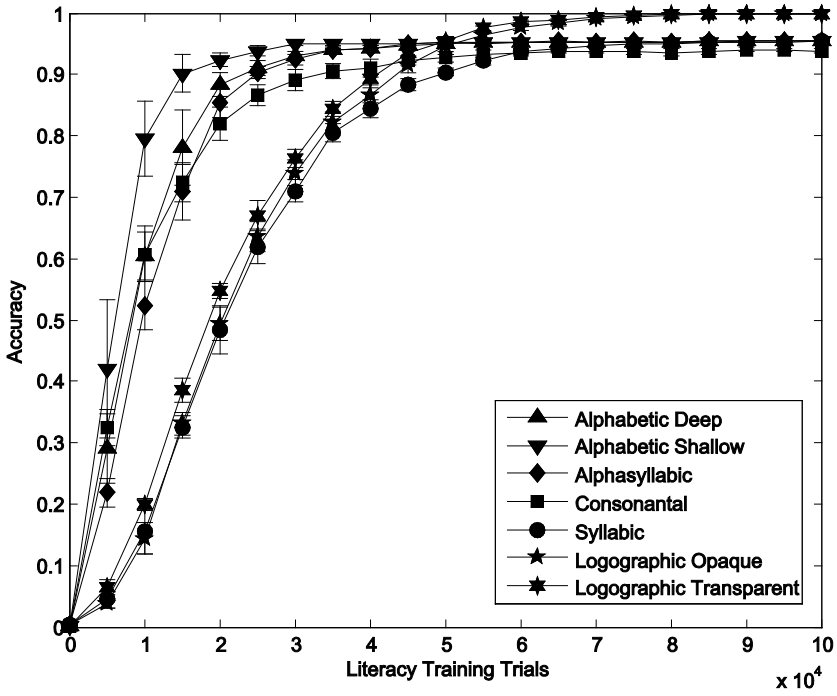

Figure 5: Reading Comprehension accuracy across training.

To examine whether systems differed in the amount of training required to achieve 90% accuracy on the reading comprehension task a one-way ANOVA was performed. 90% accuracy was selected as an appropriate threshold for comparison as it reflects a level of proficiency attainable for all networks. As can be observed in Figure 6, systems differed in this measure, $F(6,49) = 78.19$, $\eta^2 = 0.905$, $p < .001$.

In order to examine which individual orthographic systems differed in the training required to reach 90% accuracy on orthographic to semantic mappings, we compared the number of training trials required to reach this level of proficiency using six two-sample t-tests (Alphabetic Shallow, Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic

Transparent, Logographic Opaque) correcting for multiple comparisons using a Bonferroni correction. This analysis revealed that alphabetic shallow networks reached 90% accuracy prior to alphabetic deep networks (M = -8750.0, SD = 3720.1, t(14) = -4.704, p < 0.001). There was no difference in the number of trials required to reach this level of proficiency between networks trained on alphabetic deep and alphasyllabic systems (M = -2500.0, SD = 3952.8, t(14) = -1.265, p = 0.227). Consonantal systems achieved 90% accuracy prior to syllabic systems (M = -16250, SD = 5901.0, t(14) = -5.508, p < 0.001) as did logographic transparent systems (M = 7500,0, SD = 2988.1, t(14) = 5.020, p < 0.001). Differences between alphasyllabic networks and consonantal networks (M = -8125.0, SD = 5957.4, t(14) = -2.728, p = 0.016) and between logographic transparent networks and logographic opaque networks (M = -2500.0, SD = 2314.6, t(14) = -2.160, p = 0.049) were marginal suggesting faster acquisition in alphasyllabic and logographic transparent systems respectively yet these differences were not significant after correcting for multiple comparisons.

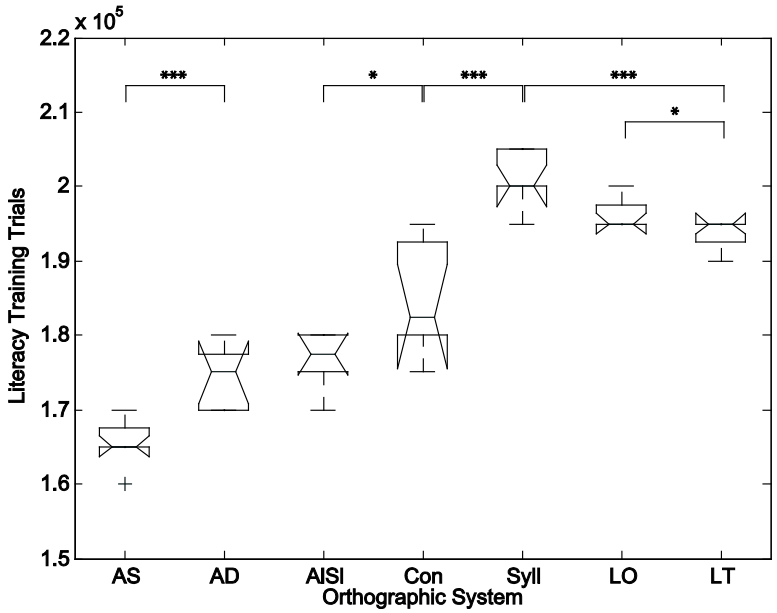

Figure 6: Literacy training trails required for network to exceed 90% on reading comprehension trials. [AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent] [2].

For acquisition of reading comprehension abilities, the level of semantic transparency embedded in the logographic transparent system had only a marginal effect on reading comprehension acquisition rate, this level of semantic transparency was also insufficient to override the advantage gained by the level of phonological transparency embedded in

alphabetic, alphasyllabic and consonantal systems in learning orthographic to semantic mappings. Our modeling shows that within such an interactive system, phonological transparency significantly increases the rate of reading comprehension acquisition. This prediction does not fit with empirical data highlighted in Seidenberg (2013), such as increased comprehension abilities in English literates over Welsh literates (Hanley, Masterson, Spencer & Evans, 2004), which suggests a more complex relation between comprehension and phonological transparency. As raised by Seidenberg (2013), our data supports the position that should these more complex relations exist they are likely to be driven by factors beyond the monosyllabic word level, such as an effect of increased spoken language exposure, or result from systems not having full phonological and semantic knowledge of a language prior to literacy training (see, e.g., Monaghan and Ellis (2010) for connectionist simulations of reading where prior experience of words is related to partial vocabulary knowledge at the point of literacy training).

Comparing learning trajectories for phonological decoding and reading comprehension

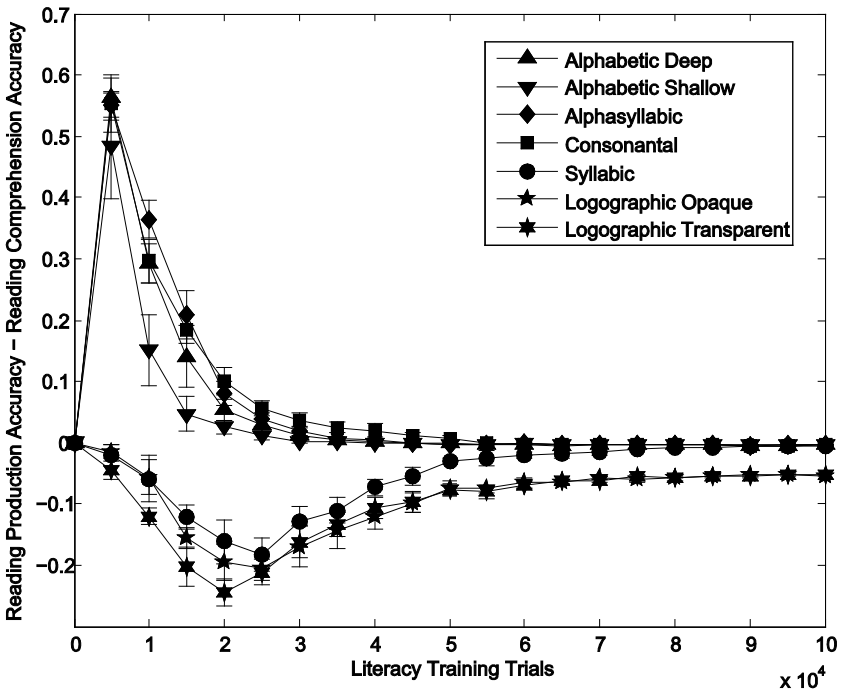

Figure 7: Difference between phonological decoding accuracy and reading comprehension accuracy across training.

Our analysis shows that transparency modulates both phonological decoding and reading comprehension acquisition rates. As an initial step to examine and raise predictions for how

transparency may affect the role of comprehension and decoding abilities in literacy acquisition we compared across systems and development the difference between the proportion of words networks were able to read for production and the proportion they were able to read for comprehension. A positive difference indicates a network is able to access the phonological form of a word from its orthography before it has learnt to access its semantic form. It is therefore likely that the system makes use of the phonological information it is able to access from the orthography to support learning of orthographic semantic mappings. Conversely, should a network display a negative production – comprehension difference, this may suggest that the system recruits semantic information it is already able to extract from a given orthographic representation to support learning of orthographic to phonological mappings.

Figure 7 presents the difference between phonological decoding and reading comprehension accuracy for each orthographic system over the course of literacy training. As systems converge towards similar levels of accuracy on both tasks come the end of training we examined whether there were differences between systems early on in training, in the first 250,000 training trials. A one-way ANOVA compared the difference between phonological decoding accuracy and reading comprehension accuracy summed over the initial 250,000 training trials (see figure 8). This revealed that systems differed in this measure $F(6,49) = 250.41$, $\eta^2 = 0.968$, $p < .001$. To examine how individual systems differed from one another systems were compared on this measure using six two-sample t-tests (Alphabetic Shallow, Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic Transparent, Logographic Opaque) correcting for multiple comparisons using a Bonferroni correction. This analysis revealed that alphabetic shallow networks displayed greater accuracy in decoding relative to comprehension than alphabetic deep networks at this early stage of training ($M = 0.359$, $SD = 0.221$, $t(14) = 3.246$, $p = 0.006$). Alphasyllabic systems also displayed greater accuracy in decoding than comprehension relative to alphabetic deep networks ($M = -0.520$, $SD = 0.217$, $t(14) = -4.786$, $p < 0.001$), although there was no difference between alphasyllabic and consonantal networks in this measure ($M = 0.051$, $SD = 0.154$, $t(14) = 0.659$, $p = 0.520$). Consonantal networks displayed substantially greater accuracy in decoding relative to comprehension than syllabic networks ($M = 1.737$, $SD = 0.163$, $t(14) = 21.28$, $p < 0.001$). Whereas, logographic transparent networks displayed greater comprehension accuracy relative to production accuracy compared to syllabic ($M = 0.277$, $SD = 0.144$, $t(14) = 3.838$,

p = 0.002) and logographic opaque networks (M = -0.194, SD = 0.098, t(14) = -3.960, p = 0.001).

This analysis reveals that logographic and syllabic systems display better performance on orthographic to semantic mappings during the initial stages of literacy training, this bias is greatest in the logographic networks, with the inclusion of semantic transparency in the orthography increasing the comprehension bias further at early stages of training.

By contrast systems that encode sub-syllabic phonological structure display a strong decoding advantage at early stages of literacy training. All such systems are able at early stages of literacy training to decode the written form of a word yet not comprehend it for a large proportion of the training corpus. This decoding advantage reduces rapidly such that all sub-syllabic phonologically transparent systems are able to both decode and comprehend written words with 90% accuracy before logographic and syllabic systems. The slot based structure of orthographic and phonological layers makes it easier for the model to identify grapheme to phoneme level correspondence. This implementation therefore ignores the additional complexity faced by readers of alphabetic languages that must not only identify the letter sound mappings, but also learn how to blend sounds together in order to produce a given word's correct pronunciation (Hudson et al., 2012). It is not possible from this data to identify the precise mechanism that drives the comprehension advantage in transparent systems. It is possible that this pattern of behaviour simply reflects the reduced complexity of orthographic to phonological mappings in such systems. Such networks are therefore likely to solve this mapping quicker and potentially with fewer resources which then allows the system to learn the more complex orthographic to semantic mappings earlier. Also compatible with the data presented in this section is the position that phonological transparency assists learning of orthography to semantic mappings due to the indirect orthography to semantics via phonology route. On the basis that networks that can map from orthography to phonology, can then use knowledge of phonology to semantic mappings to activate the semantic properties of the orthographically defined target via activation of its phonological representation. At this stage of analysis of the model it is not possible to determine the extent to which activated phonological information via orthography is recruited to assist learning of orthographic to semantic mappings. These hypotheses will be explored in further detail later in this chapter. However, the model makes a clear prediction that in transparent orthographies decoding abilities should precede comprehension and that the increased ease with which such mappings are learnt should aid reading comprehension

acquisition. As raised by Seidenberg (2013) there are few studies that directly compare phonological decoding and reading comprehension abilities and therefore whether transparency necessarily leads to increased rates of reading comprehension is still to be examined.

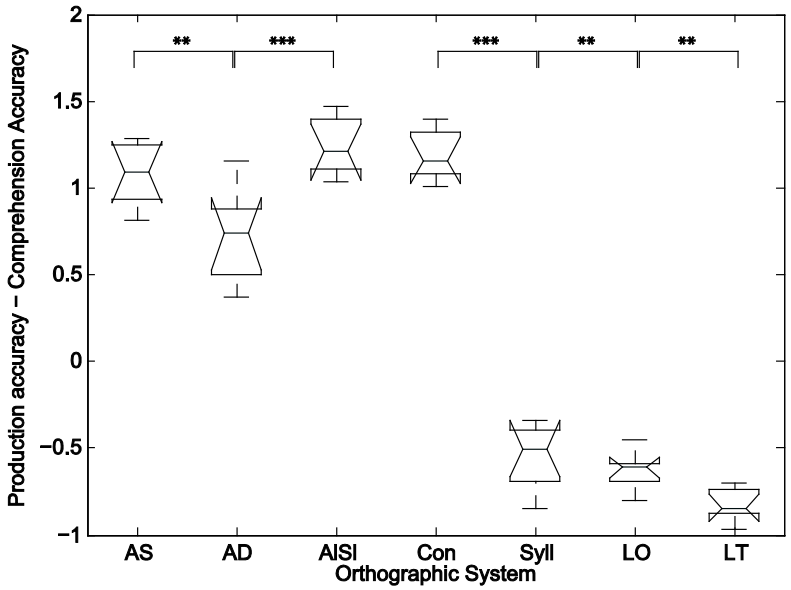

Figure 8: Summed difference between phonological decoding accuracy and reading comprehension accuracy across initial 250,000 training trials. [AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent] [2].

In contrast to the above pattern of development, logographic and syllabic systems display an early comprehension advantage. This advantage develops and recedes more gradually than the decoding advantage displayed by alphabetic, alphasyllabic and consonantal systems. Our analysis shows that semantic transparency increases the comprehension bias with logographic transparent networks displaying a greater comprehension advantage in comparison to logographic opaque networks at earlier stages of training. The fact that syllabic and both logographic systems display a comprehension bias indicates that without systematic componential relations between orthography and phonology, orthographic to phonological mappings are more difficult for networks to learn than orthographic to semantic mappings given the architecture of the implemented reading network. Within this study the hidden layer connecting orthographic and semantic layers contains twice the resources of the hidden layer connecting orthographic and phonological layers (an assumption implemented in Harm & Seidenberg, 2004). Therefore given equal complexity of mappings orthographic to semantic

mappings should be learnt quicker in this system. This highlights further complexities of deducing the mechanisms driving observed behaviour. Previous studies have argued that a greater spelling to meaning bias in Chinese compared to English is driven by the sub-lexical semantic structure embedded in Chinese orthographic representations (e.g. Yang et al., 2006). Yang et al. (2006) observed such a bias when training an implementation of the triangle model similar to that used in this paper yet trained on 103 phonological, semantic and orthographic patterns derived from Mandarin Chinese. Within our study using abstract logographic representations we replicate this bias, however as in their study such a bias may also result from differences in the resources available for learning spelling to sound or spelling to meaning mappings within the system, or inherent differences in the complexity of these mappings irrespective of the additional sub-lexical semantic systematicity.

Conclusions

The simulations demonstrate a graded effect of both phonological and semantic transparency on both the acquisition of reading for production and reading for comprehension abilities. Orthographies that encode componential sub-syllabic structure quickly learned orthography to phonology mappings while learning of orthographic to semantic mappings was delayed in such networks. However, learning orthographic to semantic mappings was still faster than systems that were not sub-syllabically transparent. This suggests that such networks may utilise phonological information they are able to extract from the orthography to aid mapping between orthography and semantics at least in early stages of training. Our data also suggests that semantic transparency may aid acquisition of reading comprehension and reading for production. In contrast to sub-syllabically transparent systems logographic and syllabic displayed greater rates of reading comprehension acquisition than reading for production, this comprehension bias being greatest in logographic semantically transparent networks. This suggests that acquisition of decoding abilities in logographic and syllabic networks may conversely be assisted by activation of semantic information via the orthography, particularly if there are systematic relationships between orthography and semantics that aid learning of such mappings.

Non-word Reading

From a computational modelling perspective, a model's ability to perform non-word reading is a critical issue as it shows that the model is able to generalise beyond the training set to novel forms and has developed sensitivity to the sub-lexical structure of orthographic to

phonological mappings. However, the close comparison of nonword reading in different orthographies provides insight into how the model represents words to support generalisation in different ways.

Alphabetic, alphasyllabic and consonantal systems were tested on their ability to read non-words (logographic and syllabic systems were not tested as these orthographic systems did not contain sub-syllabic structure). Performance was assessed once networks reached 90% accuracy on both the reading comprehension task and reading production task (to ensure comparisons were conducted at levels of proficiency attainable for all networks).

Networks were tested on their ability to read non-words using a procedure similar to that which assessed word reading detailed earlier. The orthographic input for each non-word was clamped to the orthographic layer and the phonological output examined after 12 time steps. The activation in each phoneme slot was compared to the phonological representations of all phonemes within the phoneme inventory. The phoneme whose representation was closest to the activation in each slot was recorded as the non-word produced by the model. If all phonemes produced by the model match those of the non-word whose orthographic representation was clamped to the orthographic layer then the non-word was recorded as accurately read. This differs from the procedure used to examine phonological decoding ability described earlier where the cosine distance is compared between phonological layer activity and the phonological representations of all words in the training corpus at the word level. This procedural difference allows us in the case of non-word reading to examine whether alphabetic, alphasyllabic and consonantal networks learn from the systematic phoneme grapheme correspondence at a sub-syllabic level to generalise at this level to novel forms. Logographic and syllabic systems do not possess such sub-syllabic structure and therefore can't be compared at this level.

Each category of network was tested on its ability to read 50 non-words (see section 2 for details of how non-words were constructed), with each word tested 10 times. Figure 9 displays the proportion of non-words that were read accurately (activation in all five phoneme slots was closest [cosine distance] to the corresponding phonemes within the non-word) by each category of network. All systems perform above chance and therefore this data demonstrates that alphabetic, alphasyllabic and consonantal networks were able to generalize to novel forms. A one-way ANOVA conducted on this data shows that systems differed in their performance, $F(3,28) = 138.43$, $\eta^2 = 0.937$, $p < .001$. Three two-sample t-tests were also

performed to examine differences between individual systems (alphabetic shallow, alphabetic deep; alphabetic deep, alphasyllabic; alphasyllabic, consonantal), a Bonferroni correction was applied to correct for multiple comparisons. This analysis showed that alphabetic shallow and alphabetic deep networks did not differ in their performance (M = 0.083, SD = 0.078, t(14) = 2.121, p = 0.052). Alphabetic deep networks did however perform better at reading non-words that alphasyllabic networks (M = 0.370, SD = 0.078, t(14) = 9.491, p < 0.001), while alphasyllabic networks out performed consonantal networks (M = 0.275, SD = 0.084, t(14) = 6.511, p < 0.001).

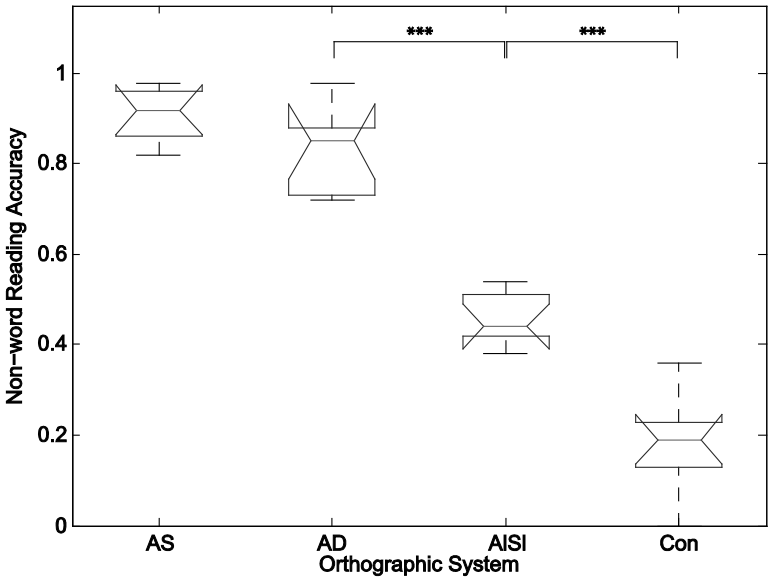

Figure 9: Accuracy of systems on non-word reading task. [AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal] [2].

To isolate the origin of differences between systems, separate one-way ANOVAs were also conducted on the accuracy of reading consonants embedded within the non-words (see Figure 10) and the accuracy of reading vowels embedded within the non-words (see Figure 11). Systems differed in their vowel reading accuracy ($F(3,28) = 184.49$, $\eta^2 = 0.952$, $p < .001$), but not in their consonant, reading accuracy ($F(3,28) = 0.650$, $\eta^2 = 0.065$, $p = 0.588$). Three two-sample t-tests were performed, with a bonferroni correction for multiple comparisons, to examine which systems differed in their vowel reading accuracy (alphabetic shallow, alphabetic deep; alphabetic deep, alphasyllabic; alphasyllabic, consonantal). This analysis revealed that alphabetic shallow systems were more accurate when reading vowels embedded in non-words than alphabetic deep networks (M = 0.087, SD = 0.048, t(14) = 3.652, p =

0.003). Alphabetic deep networks outperformed alphasyllabic networks on this measure (M = 0.310, SD = 0.088, t(14) = 7.020, p < 0.001), while alphasyllabic systems performed better than consonantal systems (M = 0.378, SD = 0.092, t(14) = 8.246, p < 0.001). Therefore systems only differed on their ability to read vowels, consonantal systems performed at chance (p = 0.2) on reading vowels within non-words (vowels are not represented in the orthography), while alphabetic shallow networks performed best followed by alphabetic deep and then alphasyllabic.

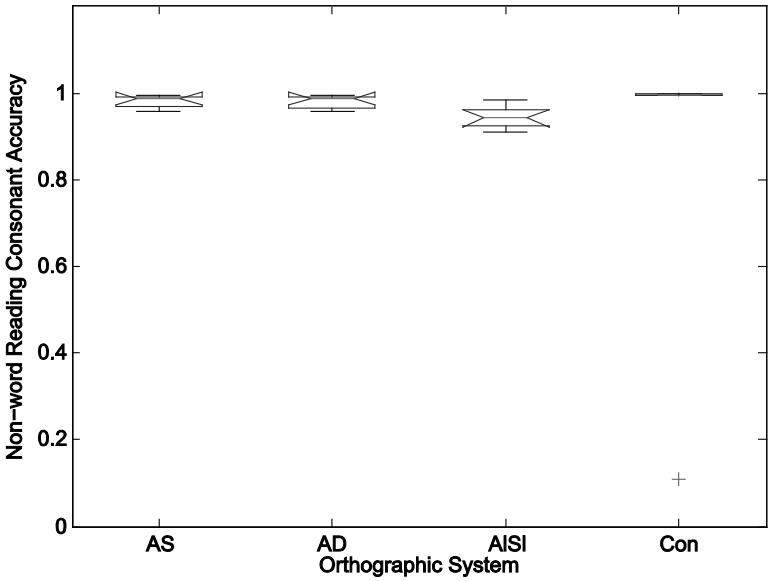

Figure 10: Accuracy of systems reading consonants embedded in non-words on non-word reading task. [AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal] [2].

Differences in non-word reading performance of networks trained on alphabetic shallow and alphabetic deep systems replicates findings within the literature. Frith et al. (1998) compared non-word reading in English and German literates at age 9 years. Comparisons between English and German are particularly insightful as they share very similar orthographic and phonological structure. Frith et al. observed that children who displayed 100% accuracy on word reading, produced 8% errors in nonword reading in the German population, whereas English children produced 22% errors. Thus, transparency appeared to increase non-word error rates. This finding was supported by Bruck, Genesee & Caravola, (1997) who attempted to control for differences in teaching methods by testing children learning to read English and children learning to read French from the same region of Canada. They also oberseved an

effect of orthographic transparency on non-word reading error rates with English children displaying 27% more errors than French children. Replicating these empirical findings alphabetic shallow networks outperformed alphabetic deep networks, figure 10 and 11 show that this difference in performance was due to greater errors made by alphabetic deep networks in mapping between graphemes and phonemes representing the vowel, as performance on consonants was similar. This data demonstrates that irregularities between orthography and phonology, which were only present in vowel mappings in the deep alphabetic system, was the factor driving reduced nonword reading performance in deep alphabetic networks. As the only point of the irregularity between letters and phonemes in the model was in the vowels, this enables us to locate a source of nonword reading differences between opaque and transparent languages being over irregularities in the mappings, and not a general processing difference across the whole word as a consequence of greater difficulty of reading opaque compared to transparent alphabetic orthographic systems.

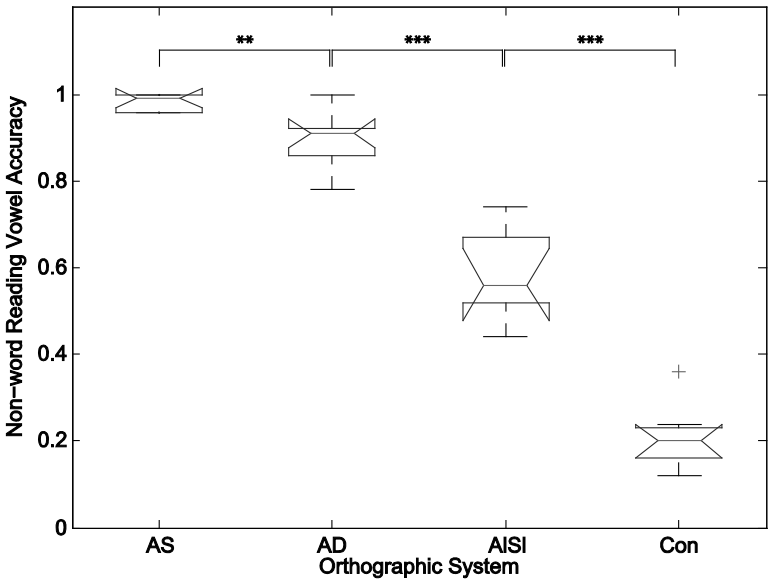

Figure 11: Accuracy of systems reading vowels embedded in non-words on non-word reading tasks. [AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal] [2].

Cross-linguistic studies comparing non-word reading across other types of orthographic system are rare. A study by Winskel & Lemwanthong (2010) showed that Thai children who by 9 years of age scored 86% on word reading tasks scored around 60% on non-word reading tasks. This might suggest that alphasyllabic literates are likely to display poorer performance

on non-word reading tasks than alphabetic literates trained on shallow orthographies. This empirical finding is consistent with the model's performance, with alphasyllabic simulations performing worse than both alphabetic systems in terms of overall accuracy, vowel accuracy and consonant accuracy.

Examining non-word reading in consonantal systems is difficult due to the under-specification of the orthography, except in cases where consonantal systems are displayed with diacritics added to represent the vowel. However, our data shows that consonantal networks were able to generalize for non-words in correctly mapping between graphemes and phonemes for consonantal units which are represented in the orthography. All errors in non-word reading for consonantal systems were due to errors made in vowel mappings, hence, these models were able to generalise even given an underspecified input representation, indicating that the model had not overlearned mappings between written and spoken word forms at the expense of processing sub-lexical mappings.

Division of Labour

We have already seen in this chapter that orthographic transparency leads to differences in reading comprehension and phonological decoding acquisition rates which suggest that transparency is likely to influence the division of labour across processing paths. Understanding the meaning of a written word can be solved either by mapping from orthography directly to semantics or via its phonological representation. Within this section we examine how the structure of the orthographic systems affects how the model solves the task of learning to read using the two routes available for each of the tasks. As previously mentioned, the model can learn to map from orthography to phonology via the direct route, or indirectly via orthography to semantics, and then semantics to phonology. Similarly, the model can learn the orthography to semantics mappings directly, or via the indirect orthography to phonology to semantics pathways. Note that the phonology to semantics and semantics to phonology pathways are already operating at a high degree of accuracy before the model is trained to learn mappings from orthography, so depending on the ease of learning the mappings from orthography to phonology and semantics can determine the extent to which the indirect pathways contribute to learning, and the degree to which they contribute changes as the reading system matures. Our analysis was interested in how each network successfully solves the reading task, therefore the following analyses includes data only for words that, at a given stage of training, the network was able to read accurately.

First, we report the model's activation of the phonological and semantic representations of each word over the 12 time steps of a reading trial, to demonstrate how gradually the model accumulated information about the word. Direct pathways provide rapid information, whereas slower activations indicate a greater role for activation contributed by the indirect pathways. Second, we directly measured the flow of information between the different pathways in the model to determine the extent to which the indirect pathways were contributing to division of labour differently for each orthographic system, both through the course of learning to read, and also, in the mature reading system, during the course of a single reading trial.

Harm & Seidenberg (2004) examined the distribution of labour within a model of English reading using a similar approach. In a fully trained model they examined the flow of activation into the semantic layers via direct and indirect paths. They observed that activation entering the semantic layer via the direct path increased most rapidly. Whereas activation entering the layer via the phonological path increased more slowly as phonological units needed to be activated first by the orthography before they could begin to exert an influence via the indirect path. Within the same paper the division of labour was also examined across development by lesioning individual paths within the network at various stages of development. They observed that in early stages of training on the deep alphabetic orthography of English the intact model and isolated phonological path achieved similar levels of performance. In contrast performance of the isolated direct semantic path improved gradually, exceeding that of the isolated phonological path at later stages of training. Both paths however were required in order to read all words within the training corpus.

Lesioning either path within networks trained in the current study lead to severely impaired performance on both comprehension and decoding tasks, with accuracy for all systems dropping to under 5%. Therefore no further description of the performance of lesioned models is reported. These lesioning results do however demonstrate a critical involvement of both paths during reading comprehension and phonological decoding for all orthographic systems. This difference between the results reported in Harm & Seidenberg, (2004) is likely due to the fact that in their study weights within the pre-literate network were fixed over the course of literacy training however, in this study literacy training trials were interleaved with pre-literacy training trials. This design decision ensured that learning within phonological and semantic networks could continue during literacy training and hence capture any effects of literacy training on semantic and/or phonological processing.

Activation of representations during reading task

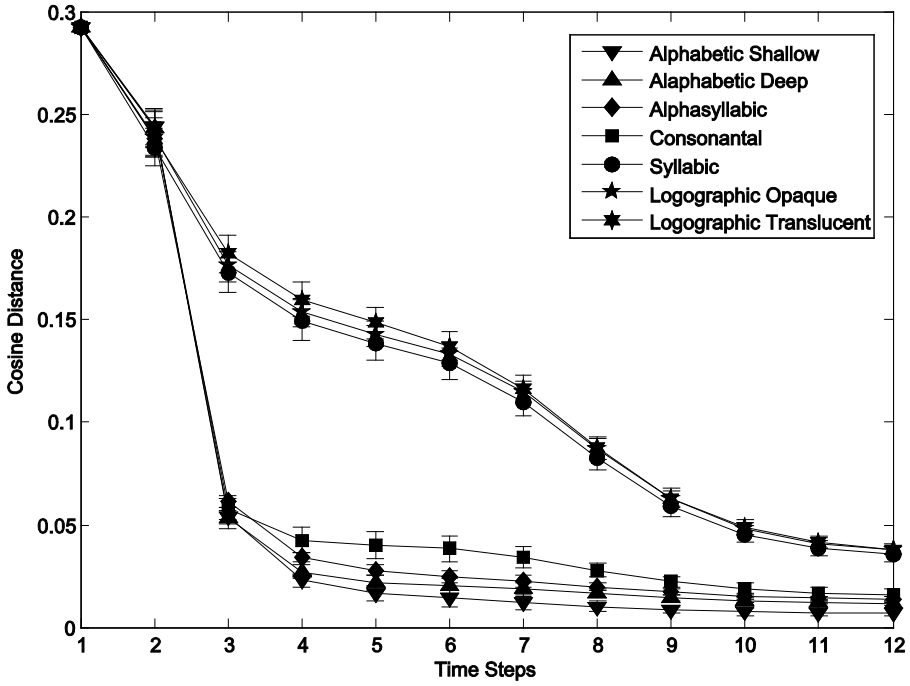

Figure 12: Distance of phonological layer activity from phonological representation of word presented to orthographic layer during reading trials in trained networks (90% accuracy on both reading comprehension and phonological decoding).

To examine differences in the timing of activation of semantic and phonological information during reading we recorded activity in phonological and semantic layers during reading trials in models that performed both reading comprehension and phonological decoding with 90% accuracy. On a reading trial a word's orthographic representation was clamped to the orthographic layer, and the activity in the semantic and phonological layers was then recorded for the following 11 time steps as the network was free to cycle. The cosine distance was calculated between activation in the phonological layer and the word's target phonological representation (see Figure 12) and the cosine distance between activation in the semantic layer and the word's semantic representation (see Figure 14) at each time step. This process was performed for all words within the corpus with figures displaying average performance calculated across all items and instantiations.

In order to understand whether systems differed in the manner in which phonological representations were activated during reading using a one-way ANOVA we compared between orthographic systems by simulation the cosine distance between phonological layer

activity and the phonological representation of the target word summed across the entire reading trial (ts 1 – 12). We corrected for differences in this measure between systems at the end of the reading trial (ts 12) by subtracting the distance at the end of the training trial from the overall mean for each system, thus ensuring results were not driven by differences between systems in the final level of activation they were able to achieve. This analysis revealed a significant difference between orthographic systems in the manner in which phonological information regarding the target item was activated ($F(6,49) = 157.98$, $\eta^2 = 0.951$, $p < 0.001$). In order to examine how individual systems differed from one another six two-sample t-tests were conducted (Alphabetic Shallow, Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic Transparent, Logographic Opaque) correcting for multiple comparisons using a Bonferroni correction. This revealed a difference between consonantal and syllabic systems ($M = -0.400$, $SD = 0.073$, $t(14) = -10.938$, $p < 0.001$), while the remaining five comparisons did not show a significant difference ($|t| < 1.5$, $p > 0.15$).

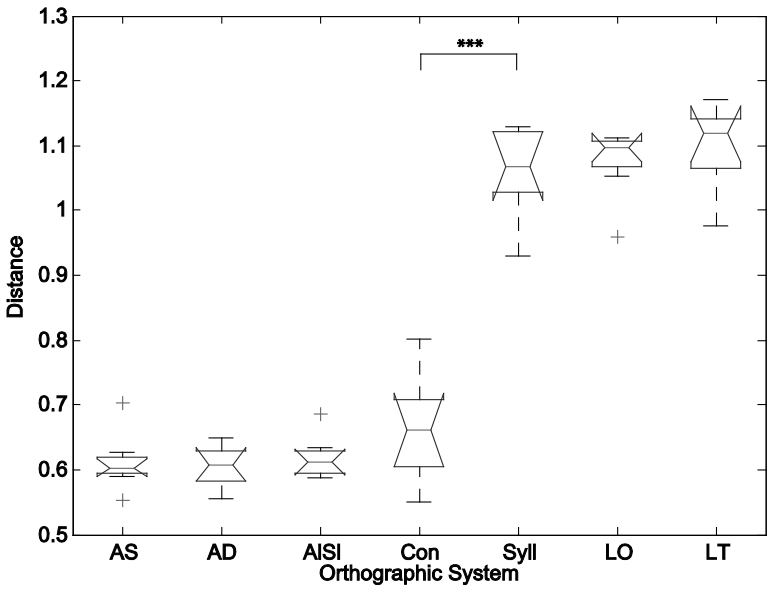

Figure 13: Mean cosine distance between phonological layer activity and phonological representation of target word presented to orthographic layer calculated across all time steps in reading trials, all items and all simulations for each orthographic system and baseline corrected for differences in cosine distance between systems at time step 12 of reading trials. [AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal;

Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent] [2].

The same analysis was performed on measures of semantic activation presented in figure 14. Figure 15 presents for each orthographic system the summed cosine distance between semantic layer activity and the semantic representation of the target word presented to the orthographic layer across the entire reading trial (ts 1 – 12) averaged over all items and simulations, and corrected for differences across systems in final levels of activation at time step 12. A one-way ANOVA performed on this measure revealed that systems differed in the manner in which semantic information was activated across reading trials ($F(6,49) = 39.52$, $\eta^2 = 0.829$, $p < 0.001$). Six two-sample t-tests (Alphabetic Shallow, Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic Transparent, Logographic Opaque) showed that consonantal and syllabic systems differed in the manner in which semantic representations were activated across reading trials ($M = 0.398$, $SD = 0.102$, $t(14) = 7.785$, $p < 0.001$), while no other comparison revealed a difference between systems.

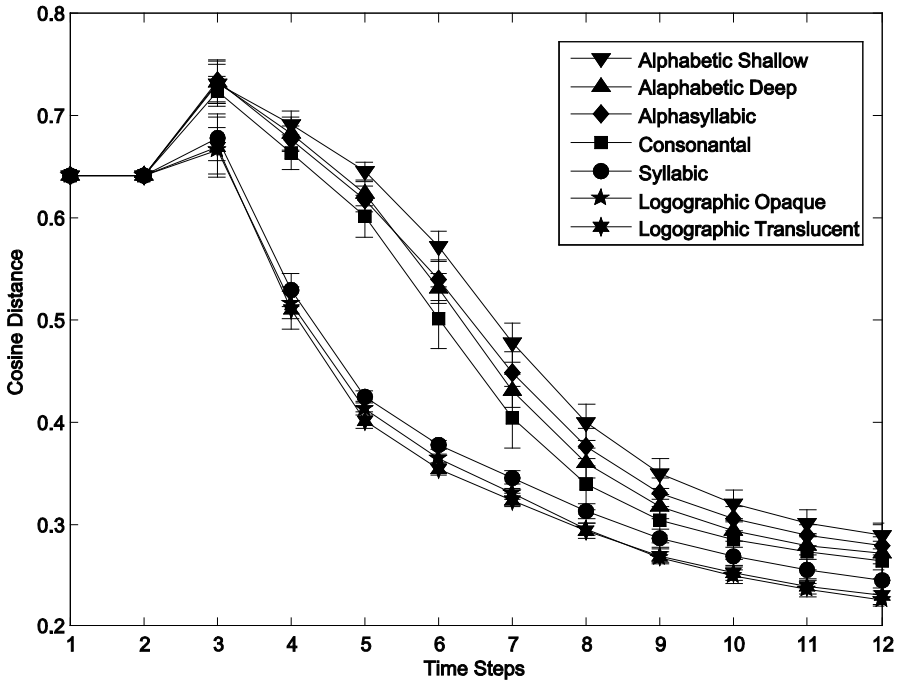

Figure 14: Distance of semantic layer activity from semantic representation of word presented to orthographic layer during reading trial in trained networks (90% accuracy on both reading comprehension and phonological decoding).

This analysis shows that alphabetic, alphasyllabic and consonantal networks rapidly activate a rich phonological representation of the target item with much of the detail of its phonological form activated by time step 3 of the reading trial. Figure 14 shows that at the same point in time activation in the semantic layer is furthest from that of the target in such networks. It is only after time step 3 that activation in the semantic layer begins to move closer to that of the targets semantic representation in alphabetic, alphasyllabic and consonantal networks.

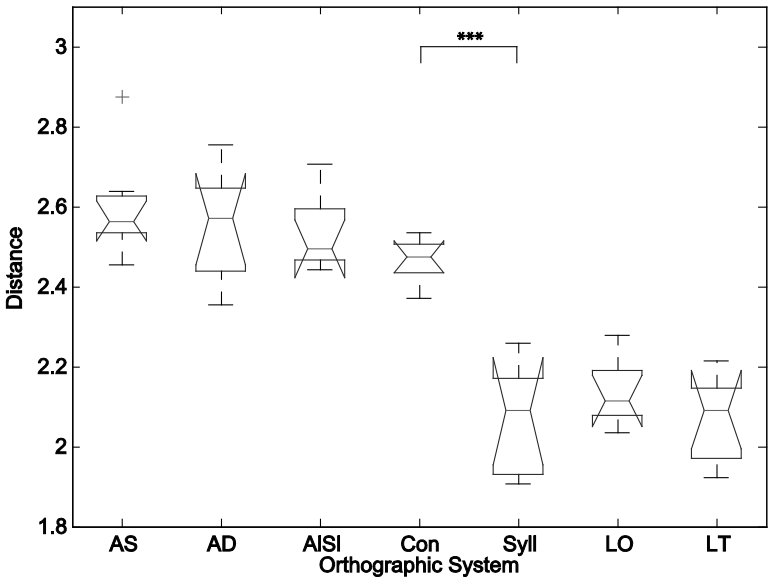

Figure 15: Mean cosine distance between semantic layer activity and semantic representation of target word presented to orthographic layer calculated across all time steps in reading trials, all items and all simulations for each orthographic system and baseline corrected for differences in cosine distance between systems at time step 12 of reading trials. [AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent] [2].

Logographic and syllabic networks however display a different pattern of behaviour. Figure 12 shows that such networks are slower to activate an items phonological form with its activation in the phonological layer steadily increasing over the course of the trial. Activation of the words semantic properties in such networks' semantic layers also increases gradually over the course of the reading trial from time step 3 onwards. However, the rate at which

semantic properties are activated is greater in syllabic and logographic networks at earlier stages of the reading trial.

The model exposes contrasts in how transparency may differentially affect processing within a universal reading architecture. Results displayed in figure 12, show that for sub-syllabic phonologically transparent systems (alphabetic, alphasyllabic and consonantal), exposure to the written form of a word activated its corresponding phonological properties rapidly. In contrast, for logographic and syllabic systems, phonological properties were activated at a slower rate over the course of the entire word reading trial. However, figure 14 shows that the semantic properties of the visually presented word are activated more rapidly in logographic and syllabic systems than alphabetic, alphasyllabic and consonantal, with phonological transparency appearing to have a graded effect on the rate of activation across transparent systems, leading to slower activation. Although, it is possible that there will be convergence across systems in the rate of activation of semantic and phonological information via orthography, the model predicts at least in early stages of literacy training differences in the relative timing of activation of semantic and phonological information should be present across systems.

Distribution of Activation

To examine how orthographic structure affects the flow of activation throughout the model during reading comprehension and phonological decoding, we recorded the flow of activation into the semantic layer and phonological layer via either the direct orthographic pathway or the indirect pathway during reading. An orthographic representation was clamped to the orthographic layer while the rest of the network was left to cycle freely for 11 further time steps. The log ratio between activation entering the semantic and phonological layer via the indirect path and activation entering the semantic layer via the direct path was calculated at each time step [log(indirect path / direct path)]. A log ratio of zero indicates the level of activation entering the layer via each path is equal, a positive value indicates increased activation via the indirect path, while a negative value indicates increased activation via the direct path. Simulations were tested on all words in the corpus and at each stage of literacy training.

Distribution of activation during reading production

Figure 16 displays the log ratio of activation entering the phonological layer via the indirect path (orthography to phonology via semantics) compared to activation entering the

phonological layer via the direct path (orthography to phonology) calculated at time step 12 of reading trials following each additional 50, 000 training patterns for each orthographic system averaged over all items and simulations. All systems display an initial negative ratio indicating that at the onset of literacy training there are greater levels of activation entering the phonological layer via the direct path than the indirect path.

In order to understand the dynamics of activation entering the phonological layer over the course of reading trial in mature networks the log ratio between activation entering the layer via the indirect path and activation entering the phonological layer via the direct path was also calculated at each time step of a reading trial once networks were able to perform both reading comprehension and phonological decoding tasks for 90% of words in the training corpus (a 90% threshold ensures all systems are able to achieve this level of performance and maturity is comparable across systems). This measure is plotted in figure 17 showing the average ratio for each system at each time step averaged across all items and simulations.

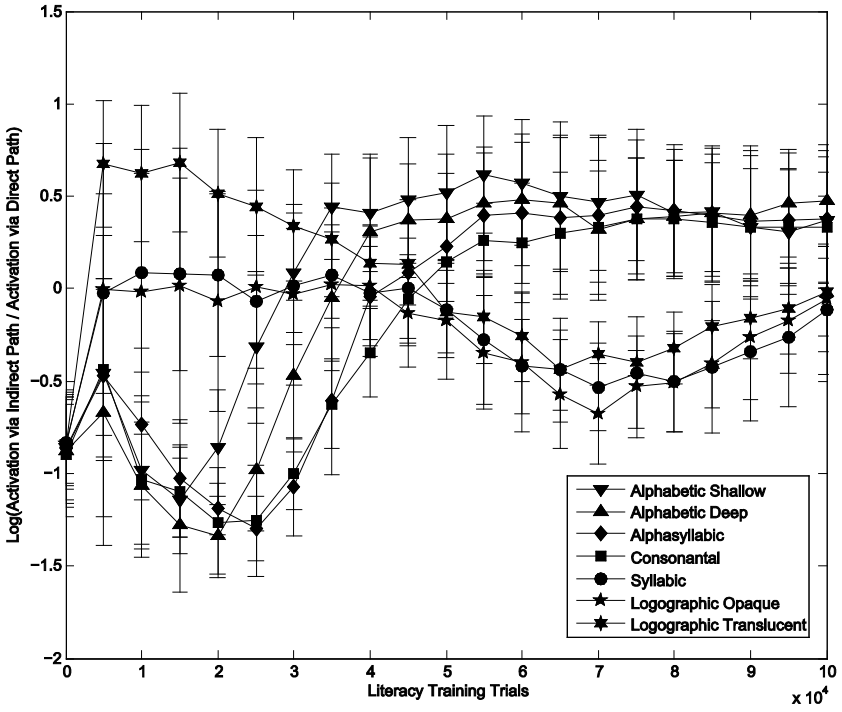

Figure 16: Log ratio of activation entering phonological layer via indirect path / via indirect path during reading trials across training averaged over all items and simulations.

To test whether differences in orthographic structure generate differences in the ratio of activation entering the phonological layer in mature systems (90% accuracy on phonological decoding and reading comprehension measures) the mean ratio across all items was summed

across all time steps for each simulation (see figure 18). Systems were then compared by simulation on this measure using a one-way ANOVA. This revealed a difference between systems in the ratio of activation entering the phonological layer via indirect and direct paths ($F(6,49) = 4.11$, $\eta^2 = 0.335$, $p = 0.002$). Six two-sample t-tests were performed to examine how individual systems differed from one another (Alphabetic Shallow, Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic Transparent, Logographic Opaque), these tests showed a less activation entering the phonological layer via the indirect path in alphasyllabic systems compared to consonantal systems ($M = -5.516$, $SD = 4.005$, $t(14) = -2.754$, $p = 0.016$), although this was no longer significant when correcting for multiple comparisons (bonferroni correction). No other comparison proved significant ($|t| < 1$, $p > 0.35$).

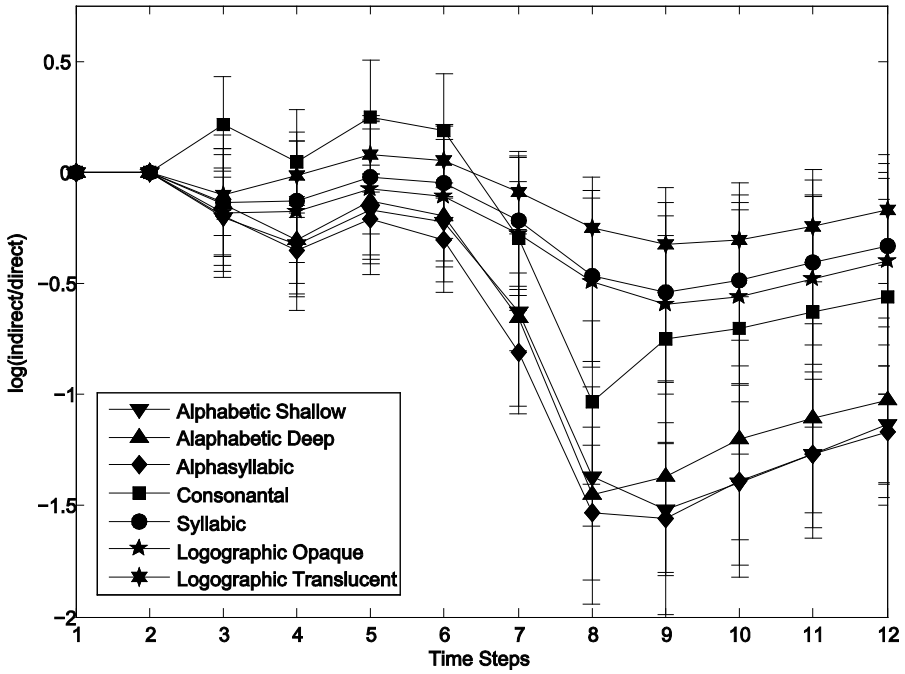

Figure 17: Log ratio of activation entering phonological layer via indirect path / via direct path over the course of a reading trial in trained (90% reading comprehension and phonological decoding accuracy) networks.

Although the analysis conducted on mature networks reveals moderate differences between systems in their use of direct and indirect paths during phonological decoding, the data presented in figure 16 suggests that systems encoding sub-syllabic structure display a

substantially different developmental trajectory in this measure to those systems that only encode structure at the syllabic level and beyond. There is a suggestion of a difference as indicated by the figure at earlier stages of training with greater indirect route bias in syllabic and logographic systems whereas at later stages of training this trend is reversed, with a suggestion of greater indirect bias in systems that encode phonological information at sub-syllabic levels.

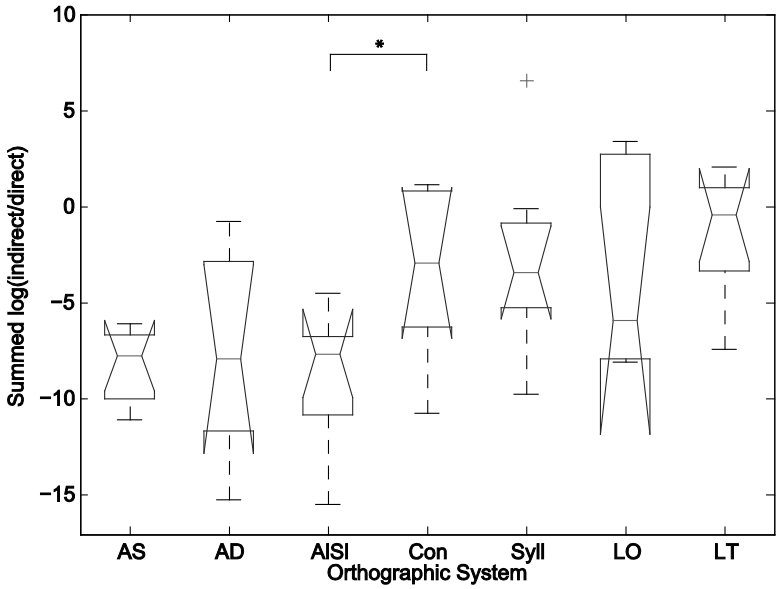

Figure 18: Log ratio of activation entering phonological layer via indirect path / via direct path at end of reading trial in trained (90% reading comprehension and phonological decoding accuracy) networks. AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent [2].

Distribution of activation during reading comprehension

To examine how orthographic structure affects the flow of activation throughout the model during reading comprehension, we performed the same analysis on measures of activation entering the semantic layer via direct (orthography to semantics) and indirect paths (orthography to semantics via phonology).

The log ratio between activation entering the semantic layer via the indirect phonological path and activation entering the semantic layer via the direct path was calculated at time step 12 of reading trials for all items and simulations following every 50,000 additional training

patterns. Figure 19 shows how this ratio averaged over all items and simulations changed over the course of literacy training for each orthographic system. The dynamics of activation entering the semantic layer via direct and indirect paths were also examined over the course of a single reading trial in networks once they were able to perform reading comprehension and phonological decoding tasks for 90% of words in the training corpus (90% threshold chosen to ensure comparisons across systems could be conducted at equivalent levels of performance), the results of which can be viewed in figure 20.

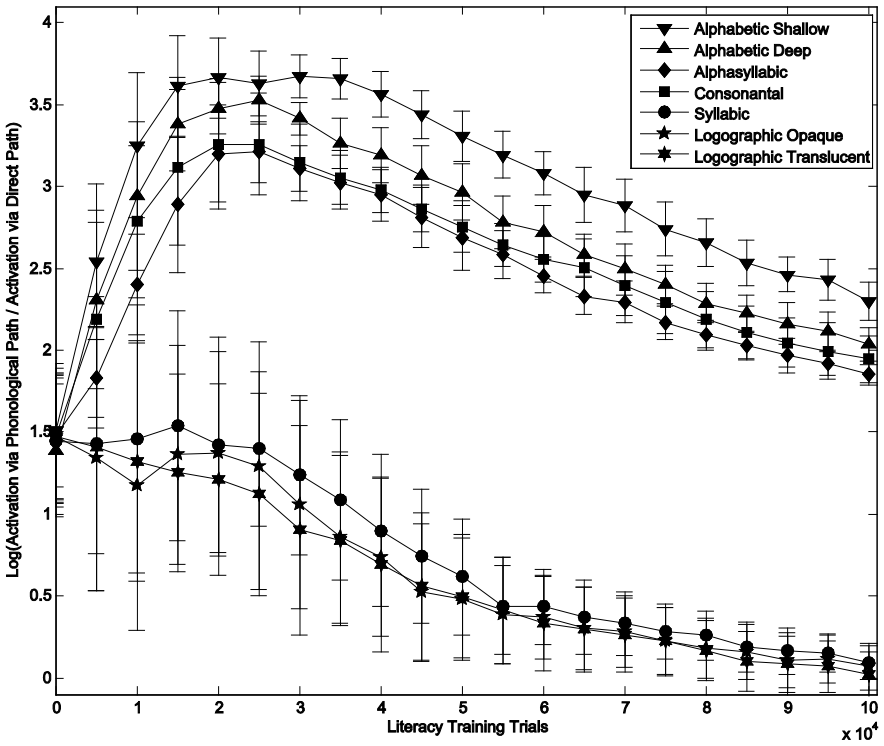

Figure 19: Log ratio of activation entering semantic layer via indirect path / via indirect path during reading trials across training averaged over all items and simulations. AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent.

To examine whether the structure of the orthographic system affected these dynamics a one-way ANOVA compared orthographic systems by simulation the log ratio of activation entering the semantic layer via the indirect path over the direct path summed across the entire reading trial (ts 1 -12) in mature networks (see figure 21). This revealed an effect of orthographic structure ($F(6,49) = 76.51$, $\eta^2 = 0.904$, $p < 0.001$). Six two-sample t-tests were

also performed to examine how individual systems differed from one another (Alphabetic Shallow, Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic Transparent, Logographic Opaque). These tests suggest that alphabetic deep networks displayed a greater bias towards activation entering the semantic layer via the indirect phonological path than alphasyllabic networks ($M = 3.765$, $SD = 2.535$, $t(14) = 2.970$, $p = 0.010$), although this difference was not significant after correcting for multiple comparisons. Consonantal networks displayed a far greater bias towards activation entering the semantic layer via the indirect path compared to the direct path when compared to ratios displayed by syllabic networks ($M = 27.371$, $SD = 4.576$, $t(14) = 11.963$, $p < 0.001$), which remained significant after correcting for multiple comparisons (bonferroni correction). Remaining comparisons did not indicate a difference between systems ($|t| < 1.2$, $p > 0.25$).

The data presented in figure 19 suggests that all orthographic systems displayed increased activation flowing into the semantic layer via the indirect path at early stages of training, yet this bias decreased over the course of literacy training. The bias displayed by all networks prior to literacy training of increased activation entering the semantic layer via the indirect path is likely to reflect the pre-established learning of phonological to semantic mappings. Given that this bias reduces from the onset of literacy training in logographic and syllabic simulations, this indicates a growing influence of activation from the direct path. Logographic and syllabic systems displayed approximately equal activation flowing into the semantic layer from each route by the end of literacy training. The bias toward greater activation via the indirect path is much greater in alphabetic, alphasyllabic and consonantal networks at all stages of development. This difference between networks emerges rapidly once literacy training commences and is greatest at earlier stages of training. In direct contrast to the pattern of development displayed by logographic and syllabic systems, alphabetic, alphasyllabic and consonantal networks display an immediate and rapid increase in indirect bias from the onset of literacy training. This bias peaks after approximately 30,000 literacy training trials, before decreasing for the remainder of training. This pattern indicates an increasing influence of the indirect path at early stages of literacy training, with the direct path becoming more influential as training increases after this point. By the end of training similar to earlier stages of training, alphabetic, alphasyllabic and consonantal networks continue to display a far greater indirect bias compared to logographic and syllabic networks. This contrasts with some arguments in the literature that suggest the indirect path is likely to

become largely redundant in proficient readers (see Levy 2008; 2009) However, it is not possible to rule out this result arising should the model by exposed to extensive literacy training with pressure to activate semantic knowledge rapidly therefore favouring direct path activation.

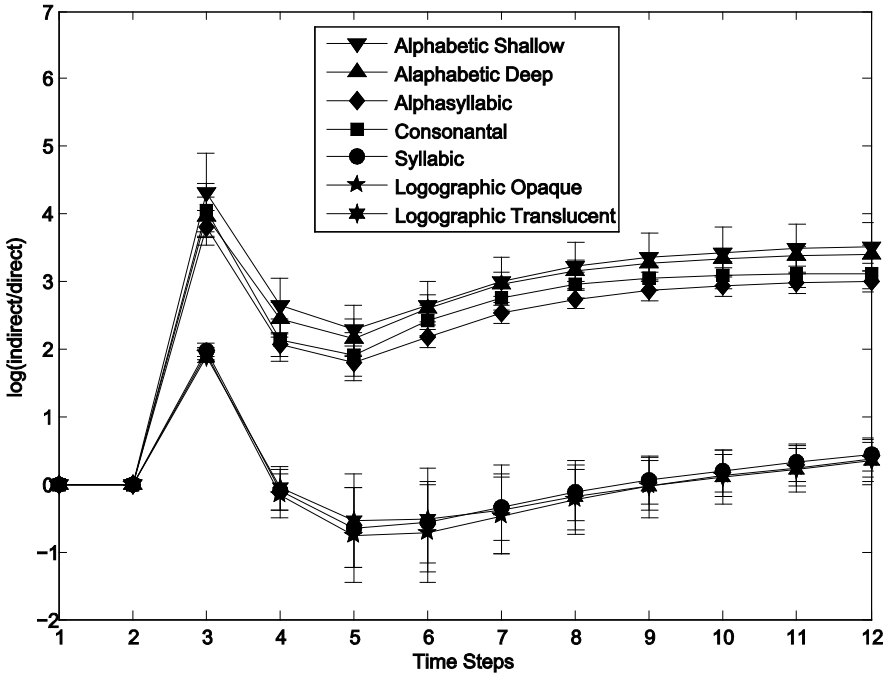

Figure 20: Log ratio of activation entering semantic layer via indirect path / via direct path over the course of a reading trial in trained (90% reading comprehension and phonological decoding accuracy) networks .

Figure 20 shows that the ratio of activation entering the semantic layer via the indirect path and direct path varies over the course of a single reading comprehension trial. All orthographic systems displayed an initial increase in activation into the semantic layer via the indirect path early in the reading trial. In logographic and syllabic simulations this increased activation via the indirect path declined rapidly and remained low for the remainder of the trial. In contrast alphabetic, alphasyllabic and consonantal simulations displayed a greater initial increase in activation via the indirect path. This also declined rapidly, however in contrast to syllabic and logographic simulations the indirect path contribution remains greater in such networks for the remainder of the trial and increases steadily from time step 3 onwards.

All networks display an initial spike in activation from the phonological path at time step 3, although this spike is far larger in sub-syllabic transparent networks. This coincides with activity in the semantic layer becoming more distant from the target (figure 14), suggesting that this initial spike may record largely noise coming from the phonological layer before it begins to settle on the word's phonological form. Following the initial spike in activation from the indirect path logographic and syllabic networks display an approximately equivalent level of activation entering the semantic layer from both direct and indirect paths. Although mean ratios indicate a small bias toward direct path activation at earlier stages of the trial, activation from the indirect path steadily increases over the course of the trial such that activation through both paths are equivalent by the end of the reading trial. Sub-syllabic transparent systems display a similar trough in indirect bias at time step 5 with the indirect bias increasing for the remainder of the reading trial.

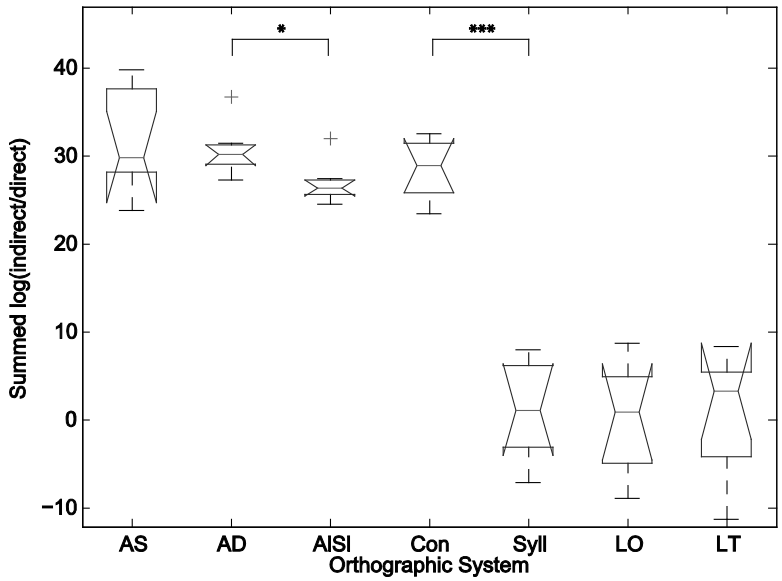

Figure 21: Log ratio of activation entering phonological layer via indirect path / via direct path at end of reading trial in trained (90% reading comprehension and phonological decoding accuracy) networks. AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent [2].

Richardson et al. (2011) observed activation of the indirect path at both early and late stages of word reading, it is possible that this is captured in our model by the two peaks we observe in indirect path activation. Richardson et al. (2011) interpreted this activation as potentially

reflecting mapping at multiple grain sizes, however, given that we observe similar peaks in activation for both transparent and non-transparent systems, so for some systems finer grain-size than the whole word would not assist the mapping between orthography and phonology, this explanation would not account for this pattern in our computational model. Instead, the initial peak is due to initial noise coming from the phonological layer prior to it settling on the phonological representation of the target, the later peak emerges as a consequence of increasing activation from the phonological layer as an increased number of phonological features of the target word are activated.

Discussion: Distribution of activation during reading

Results offer a description of how the flow of activation throughout the reading system may be affected by orthographic transparency both over the course of development and in mature systems. Analysis of ratios capturing differences in activation of indirect and direct paths in mature networks during phonological decoding shows that alphabetic and alphasyllabic systems display greater levels of activation passing through the direct orthography to phonology path than indirect semantic path in comparison to syllabic and logographic systems, although, as indicated by figure 16, this difference may alter if networks are exposed to further literacy training. By contrast a more stable difference over development in use of direct and indirect paths is displayed during reading comprehension between systems that encode sub-syllabic structure in their orthography and those that do not. Mature logographic and syllabic systems display greater activation via direct paths than indirect paths during reading comprehension in comparison to systems that encode phonological information at sub-syllabic levels, further as suggested by figure 19, this difference emerges early in development and remains present over the course of development. In contrast to predictions, there was no significant influence of semantic transparency on the distribution of activation across direct and indirect paths during reading comprehension or phonological decoding.

Together these data argue for a graded effect of phonological transparency on the distribution of activation across indirect and direct paths during both phonological decoding and reading comprehension. This aligns with cognitive neuroscientific studies comparing activation through ventral (direct) and dorsal (indirect) paths across populations that differ in the orthographic transparency of the system on which they were trained. As captured by the model, Paulesu et al. (2000) observed a dorsal bias in activation in a shallow orthographic system (Italian) compared to a deeper orthographic system (English). Further, Kiyosawa et

al., (1995) observed a dorsal bias for individuals when they read a transparent orthography (Kana) compared to a non-transparent orthography (Kanji).

As the model used in this study was a learning model it was also possible to examine how the division of labour in networks developed over the course of literacy training. Previous neuroimaging studies (Pugh et al., 2000; Shaywitz et al., 2002) and computational modelling studies (Harm & Seidenberg, 2004) have demonstrated increased activation of the indirect path during reading comprehension in early stages of literacy training on a deep alphabetic system such as English. Our modelling replicates this finding and indicates that phonological transparency leads to distinct patterns of development. By contrast semantic transparency (at the level implemented in this study) appears to have little impact on the development of the distribution of labour.

Importantly within this study the division of labour is not predefined but instead develops as the statistical learning algorithm applied attempts to find the most efficient means of mapping between orthography and semantics given the constraints imposed by the architecture and learning environment. As orthographic structure is the only factor to differ between simulations we can be confident that differences in the division of labour observed between simulations are driven by this variable and not semantic or phonological structure, factors which has not been controlled in previous modelling studies (Yang et al., 2013). The above data therefore provides an explicit description of how differences in processing observed in neuroimaging studies may emerge as a consequence of the same underlying architecture and learning mechanisms configuring around orthographic systems that differ in their semantic and phonological transparency.

Literacy effects

There is existing behavioural (e.g. De Gelder & Vroomen, 1992) and neural (e.g. Brennan et al., 2013) data that suggest that phonological processing is affected differentially by orthographic transparency, with transparency leading to finer grain processing. In this section we therefore examine whether differences in the structure of phonological representations can be observed across orthographic systems, this is possible as phonological representations are controlled across orthographic systems. Further, as semantic representations are also controlled across orthographic systems we examine whether orthographic transparency in both phonological and semantic dimensions affects the structure of emergent semantic representations. Finally, as our implementation of a logographic opaque system ensures

correspondence between orthography and phonology and orthography and semantics only exists at the word level we are able to examine in isolation the effect of orthographic similarity on emergent phonological and semantic representations.

In the following sections we first examine effects on the structure of representations activated during trials in which orthographic representations are also directly activated. We then move on to examining effects on representations during phonological and semantic retention trials in which orthographic representations are not active.

Measures were recorded only for words networks were able to read accurately in terms of activating the correct phonology and semantic representations when presented with their corresponding orthographic form, thus ensuring that our results only reflect changes to the representation of words for which the network possesses functional knowledge.

To ensure systems were compared at equivalent levels of reading proficiency networks were tests once they were able to map from orthography to phonology and semantics accurately for 90% of words in the training corpus. A threshold of 90% was chosen as this was a level of performance attainable for all systems.

Effects on phonological processing

To examine whether the orthographic system on which the model received training altered the componential nature of phonological processing, we determined the extent to which similarities between words were best reflected at the phoneme level (thus reflecting phoneme-level segmentation) or whether similarities were just in terms of global similarity across the whole word, regardless of phoneme-level similarities. This was done by testing the model on a set of 1200 words that were controlled to be similar in terms of the number of phonological features that they shared over the whole word, but differed in terms of the number of phonemes that they shared. In this set there were 400 control words (Control set), a set of 400 words with overlapping phonemes with the control set (Phoneme set), and a set with an equal number of phonological features overlapping with the control set but with fewer overlapping phonemes than the phoneme-set (Feature set). The controls were yoked, such that a single word in both the Feature set and Phoneme set was paired with a word within the Control set, these words shared the same number of phonological features as the Control set word (M = 35.25, SD = 2.39) yet the word within the phoneme set shared an increased number of phonemes (Phoneme set shared phonemes: M = 2.04, SD = 0.20; Feature set shared phonemes: M = 0.82, SD = 0.45).

Simulations were tested on their processing of these three sets of words during both reading and phonological retention tasks. Activation in phonological layers was recorded while processing each word on each task. For each word pair (i.e. Control set_i Feature set_i; Control set_i Phoneme set_i) the cosine distance [1-cosine angle between vectors] between activation in the phonological layer was calculated. The following results report the ratio [Pdist/Fdist] of these two distances (i.e. Control set to Phoneme set Distance/Control set to Feature set distance). A Pdist/Fdist lower than 1 indicates that phonological representations with shared phonemes are more similar than those that only share the same number of phonological features, indicating phoneme-level processing. A ratio at 1 indicated that only global similarity between phonological forms of words affected processing.

Granularity of Phonological Processing During Reading

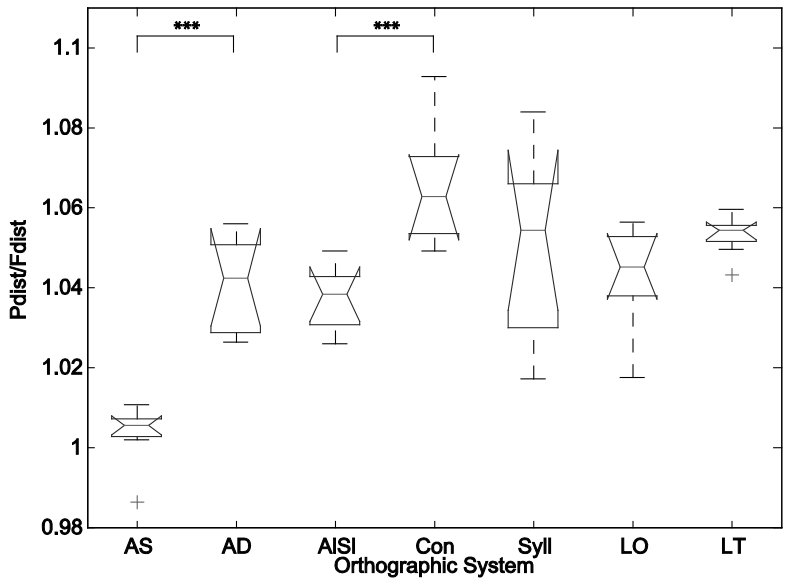

Figure 22: Pdist/Fdist displayed by trained networks (90% accuracy on reading comprehension task) recorded at ts 11 of reading task. AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent [2].

Figure 22 displays the Pdist/Fdist ratio based on activation in the phonological layer at time step 12 of reading tasks displayed by networks able to accurately map from orthographic representations to phonological and semantic representations for 90% of words within the training corpus. A one-way ANOVA conducted on this measure showed that systems differed

in Pdist/Fdist ratios at time step 12 of reading tasks, $F(6,49) = 16.92$, $\eta^2 = 0.674$, $p < .001$. Six two-sample t-tests (corrected for multiple comparisons using a bonferroni correction) examined differences between individual orthographic systems (Alphabetic Shallow, Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic Transparent, Logographic Opaque). This analysis revealed a lower Pdist/Fdist ration for alphabetic shallow networks compared to alphabetic deep ($M = -0.037$, $SD = 0.010$, $t(14) = -7.581$, $p < 0.001$), and for alphasyllabic networks compared to consonantal networks ($M = -0.028$, $SD = 0.011$, $t(14) = -4.627$, $p < 0.001$). No other comparison proved significant ($|t| < 2.02$, $p > 0.05$).

Literacy Effects on Spoken Language Processing

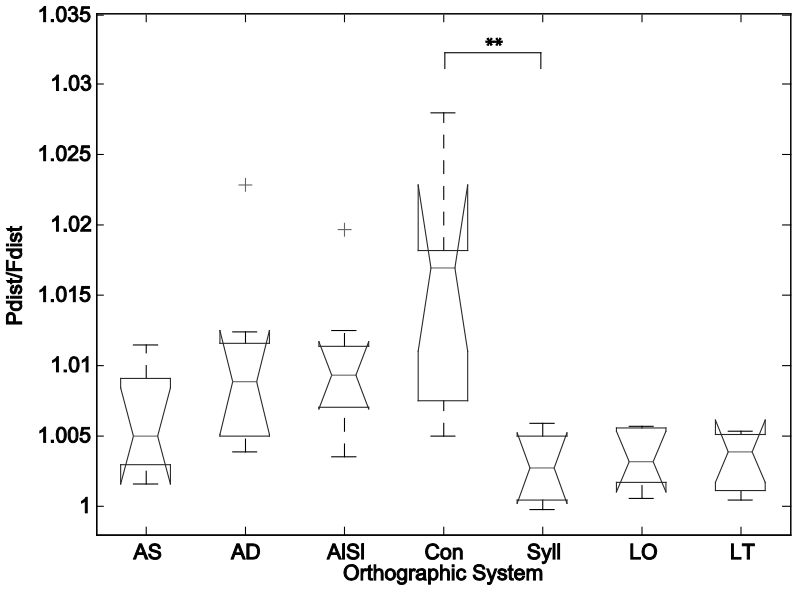

Figure 23: Pdist/Fdist displayed by trained simulations (90% accuracy on reading comprehension and phonological decoding tasks) recorded at ts 11 of phonological retention task. AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent [2].

Pdist/Fdist ratios were also examined at time step 12 of phonological retention trials performed by networks able to map from orthography to phonology and semantics accurately for 90% of words in the training corpus. Figure 23 presents the Pdist/Fdist ratio displayed by each system on retention tasks averaged across all items and simulations. A one-way

ANOVA compared systems by simulation on this ratio revealing that systems differed in this measure, $F(6,49) = 7.64$, $\eta^2 = 0.485$, $p < .001$. Six two-sample t-test (with a bonferroni correction applied for multiple comparisons) were also performed to test differences between individual orthographic systems (Alphabetic Shallow, Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic Transparent, Logographic Opaque). These tests revealed that consonantal networks displayed a greater Pdist/Fdist ratio than syllabic networks ($M = 0.012$, $SD = 0.006$, $t(14) = 4.194$, $p < 0.001$), while no other test revealed a significant difference between systems ($|t| < 1.55$, $p > 0.15$).

Discussion: Phonological Effects

Cognitive neuroscientific evidence (see section 1) suggests that effects of literacy on phonological processing may be observed as a consequence of a restructuring of phonological processing regions used for spoken word processing or due to online activation of orthographic representations during speech processing. Results reported in earlier sections of this chapter demonstrate that the orthography on which the reading system is trained is likely to determine the nature of the effect on processing. Within the current study although systems displayed identical pre-literate phonological processing behaviour, differences emerged in processing as a consequence of literacy training. Literacy training affected both the structure of the phonological representation of a word when activated by its orthographic form (online activation) and the structure of phonological representations activated in the absence of orthographic activation (phonological restructuring).

The Pdist/Fdist ratio reported above permits an examination of the extent to which systems displayed componential phonemic processing. A ratio lower than 1 indicates that items that share phonemes are processed more similarly than items that simply share the same number of phonological features. We had predicted based on the existing behavioural evidence, theoretical (Psycholinguistic Grain Size Theory) and computational models (see section 1) that orthographic training should affect phonological processing such that processing reflects the systematic relations between orthography and phonology within the given orthographic system. Therefore, shallow alphabetic systems, in which there is one to one correspondence between graphemes and phonemes, should develop stronger componential phonemic level processing than logographic systems in which correspondence between orthography and phonology only exists at the word level. Therefore transparent systems should display lower

Pdist/Fdist ratios. Literacy training however did not lead to greater similarity between items that shared phonemes for any system examined. Instead all systems displayed an increased difference as a consequence of literacy training between items that shared phonemes in comparison to items that shared an equal number of phonological features.

Examining phonological processing when a word's orthographic representation was also active revealed a graded effect of transparency on Pdist/Fdist ratios. This fits with our predictions of coarser grained processing in less transparent systems. Over the course of training on all orthographies, there is increasing phoneme-level processing, which is a consequence of recognition of phonemes comprising the spoken forms of words. However, the extent to which this affects processing is modulated by the reflection of this phoneme-level granularity in the orthographic system. This is reflected in behavioural data that shows there are early and relatively stable effects of phoneme awareness in literacy training (Alcock et al., 2010; De Jong & Van der Leij, 1999; Hulme, Snowling, Caravolas & Carroll, 2005; Treiman & Zukowski, 1991).

The simulations also reveal differences in phonological representations activated during phonological retention tasks. This analysis assessed the extent to which phonological processing was effected in the absence of orthographic activation. The restructuring effects observed in this analysis are consistent with observed orthographic consistency effects reported in the literature that phonological processing regions become restructured as a result of literacy training (e.g. Perre et al., 2009) and therefore display effects of orthographic knowledge in the absence of activation of orthographic representations. This data therefore offers an explicit description of the mechanism that may be driving these effects.

In contrast to our predictions, logographic and syllabic systems displayed little effect of literacy on phonological processing in the absence of orthographic activation. In such systems increased phoneme overlap had little effect on the similarity of representations. These data suggest that should restructuring occur within logographic or syllabic systems it has little effect on phoneme level processing in the absence of orthographic activation.

Although, the simulations and analysis detailed above predict an effect of literacy on phonological processing that is determined by the transparency of the orthographic system, a number of the models predictions are at odds with existing theoretical models and empirical findings. Specifically, increased training on orthographic systems that encode fine grained phoneme-level structure does not lead to increased fine grained phoneme level phonological

processing. A potential reason for the mismatch between model behaviour and empirical findings are the assumptions underlying the representation of speech and pre-literate phonological processing within the model. The current model assumes that pre-literate phonological representations are established through learning to retain in memory a speech signal in which phonemes are represented as discrete units, with no pronunciation variation within phoneme categories and no effect of phonological context. It is likely that this serves as an inadequate description of the speech signal for investigating emergent phonological structure, because speech is inherently noisy and inherent within natural speech are features such as co-articulation, elision and reductions (e.g. Krakow, 1999; Browman & Goldstein, 1989). Such assumptions regarding the input (speech) to the system will have profound implications for the preliterate (and post literate) phonological representations the network develops (we return to this issue in the General Discussion).

Literacy Effects on Speech Comprehension

To examine the effects of orthographic structure on semantic processing, semantic layer activity was recorded during reading and semantic retention tasks. The cosine distance between semantic layer activity, recorded when processing words of the same semantic category and activity recorded when processing words of a different semantic category was calculated. The ratio (*CatRat*) between these measures indicates the strength of the effect of semantic category on processing. A smaller *CatRat* value indicates that the representations of words from the same semantic category is closer than words from different semantic categories, providing an analogue to the phonological granularity measure, by determining the extent to which similarity among semantic representations is emphasised by the different orthographic systems. Again, analogously to the phonological structure results, we predicted that reflecting semantics in the orthography, as in the logographic transparent language, would result in enhancement of similarity in the semantic structure.

During Reading

Data in figure 24 describes how *CatRat* varies across orthographic systems at time step 12 of reading trials for networks able to map from orthography to semantics and phonology for 90% of words in the corpus. A one-way ANOVA performed on this data revealed a difference as a consequence of orthographic system ($F(6,21) = 10.21$, $\eta^2 = 0.556$, $p < 0.001$). Six two-sample t-tests (with a bonferroni correction applied for multiple comparisons) examined differences between individual orthographic systems (Alphabetic Shallow,

Alphabetic Deep; Alphabetic Deep, Alphasyllabic; Alphasyllabic, Consonantal; Consonantal, Syllabic; Syllabic, Logographic Transparent, Logographic Transparent, Logographic Opaque). This revealed a greater *CatRat* for syllabic systems compared to logographic transparent systems (M = 0.037, SD = 0.020, t(14) = 3.780, p = 0.002). Differences between logographic transparent and logographic opaque systems indicate a marginally higher *CatRat* for logographic opaque systems although this was significant after correcting for multiple comparisons (M = -0.026, SD = 0.021, t(14) = -2.492, p = 0.026). None of the remaining four tests distinguished between systems (|t| < 2.0, p > 0.05).

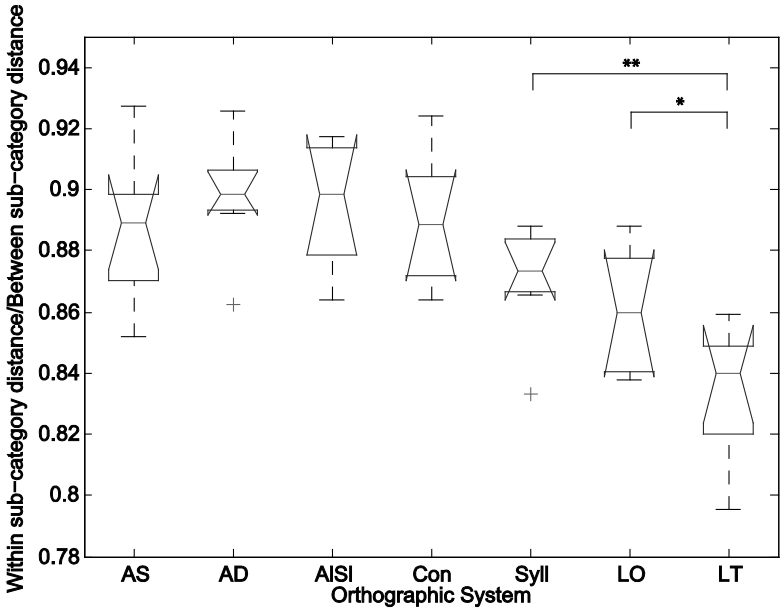

Figure 24: Ratio of distance between semantic representations within the same semantic sub-category / distance between semantic representations of different semantic categories displayed at ts = 11 of reading task in trained networks (90% accuracy on reading comprehension and phonological decoding tasks). AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent [2].

During Semantic Retention Task

CatRat was also calculated during semantic retention trials to examine effects on semantic processing during non-reading tasks when orthographic information was not active. Figure 25 displays CatRat calculated for each system, averaged over items and simulations, recorded at ts = 12 of semantic retention trials. A one-way ANOVA conducted on these ratios tested by

simulation whether systems differed in this measure. This analysis indicated that there was no difference between systems ($F(6,49) = 0.900$, $\eta^2 = 0.100$, $p = 0.500$).

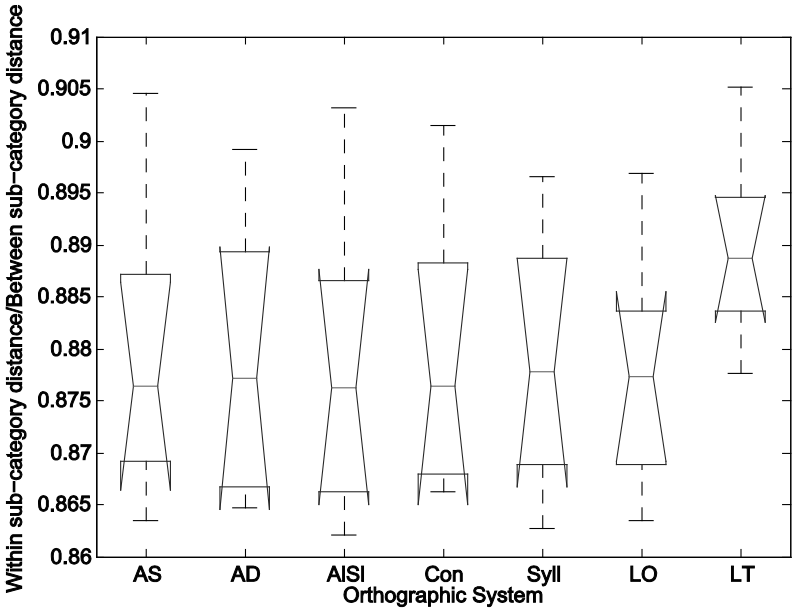

Figure 25: Ratio of distance between semantic representations within the same semantic sub-category / distance between semantic representations of different semantic categories displayed at ts = 12 of semantic retention tasks in trained networks (90% accuracy on reading comprehension and phonological decoding tasks). AS = alphabetic shallow; AD = alphabetic deep; AlSl = alphasyllabic; Con = consonantal; Syll = syllabic; LO = logographic semantically opaque; LT = logographic semantically transparent [2].

Discussion: Semantic Effects

Lower *CatRat* values indicate a greater between semantic category distinction. Our data demonstrates that in the context of orthographic activation semantic category distinctiveness can be modulated by the orthographic system, with logographic transparent systems displaying greater between category definition in semantic layer activity when corresponding orthographic representations are active (reading task), the single system that explicitly encoded semantic structure in its orthography (logographic semantically transparent). It is possible that this effect purely reflects richer semantic activation as a result of improved learning of orthographic semantic mappings however we believe this to be unlikely given that activation of semantic representations by logographic networks appear more distant from their true form (see time step 12, figure 12). Instead it seems that logographic systems

increase the likelihood of developing sensitivity to semantic category distinctions due to greater reliance on word level orthographic semantic mappings.

In contrast to the effects observed when orthographic representations are active, no significant effect was observed on levels of semantic category distinction between systems in the absence of orthographic activation, although, there is a suggestion of a difference between logographic transparent systems and none semantically transparent systems as indicated by figure 25. It may be possible to observe an effect should power be increased with further simulations or should semantic structure be more heavily encoded in the orthography. However, in contrast to predictions the numerical difference suggests that in the absence of orthographic activation semantic transparency may reduce between category distinctions.

Effects of visual similarity

Results reported earlier in this chapter demonstrate that phonological and semantic representations can be influenced by sub-word level phonological and/or semantic structure embedded in the orthography. The logographic opaque system is the only system not to encode semantic or phonological structure below the word level. It thus offers a means of examining in isolation whether visual similarity in the structure of orthographic representations affects emergent semantic and/or phonological representations.

To test this we examined semantic and phonological processing when logographic opaque networks processed words that were controlled for their level of phonological and semantic similarity yet were either visually similar or dissimilar in their orthographic representation. Three sets of 250 words were identified a control set, a set that contained words visually similar to the control set and a set that contained words visually dissimilar from the control set . The controls were yoked, such that a single word in both the visually similar set and visually dissimilar set was paired with a word within the control set, these words were controlled for overlap in both phonological and semantic dimensions yet words in the visually similar set shared an increase number of orthographic features with the words in the control set (Cosine distance between visually similar set and control set [phonological representations: M = 0.369, SD = 0.087; semantic representations: M = 0.898, SD = 0.081; orthographic representations: M = 0.706, SD = 0.027]; Cosine distance between visually dissimilar set and control set [phonological representations: M = 0.373, SD = 0.089; semantic representations: M = 0.897 , SD = 0.083; orthographic representations: M = 0.294, SD = 0.027]).

Phonological and semantic layer activity was recorded in logographic opaque networks at time step 12 of reading, semantic retention and phonological retention trials when processing each word within these three sets at two stages of training. A stage prior to literacy training but at which networks were able to perform semantic retention, phonological retention, speech comprehension and speech production tasks accurately for all words in the corpus, and a stage late on in literacy training at which networks were able to map from orthography to phonology and semantics for 90% of words within the training corpus (controls ensured networks were able to read all 750 words within the control, visually similar and visually dissimilar test sets).

At both stages of training for each task the cosine distance between phonological and semantic layer activity when processing control words and activity when processing visually similar words was calculated as was the distance between activity for control words and activity for visually dissimilar words. The ratio of the distance between the control set and visually similar set compared to between the control set and visually dissimilar set was calculated for measures extracted pre-literacy and post-literacy. Subtracting the pre-literacy ratio from the post literacy ratio provides a measure of the change in representational distance as a consequence of literacy training (*VisEffect*). Should this measure be greater than zero it indicates that orthographically similar items become more similar in either semantic or phonological dimensions as a result of literacy training, conversely a value below zero indicates representations become more distinct, whereas a value of zero indicates no change as a consequence of literacy training.

During Reading

Figure 26 displays the *VisEffect* for semantic layer activity displayed during reading tasks and for phonological layer activity displayed during reading tasks averaged over all items and simulations. To test whether orthographic similarity effected the structure of semantic and phonological representations we used two one-sample t-tests to test whether each *VisEffect* measure differed from zero. This analysis revealed that as a consequence of literacy training the distance between semantic representations whose corresponding orthographic representations were similar increased ($M = 0.092$, $SD = 0.012$, $t(14) = 21.520$, $p < 0.001$). Similarly as a consequence of literacy training the distance between phonological representations whose corresponding orthographic representations were similar also increased ($M = 0.048$, $SD = 0.018$, $t(14) = 7.641$, $p < 0.001$).

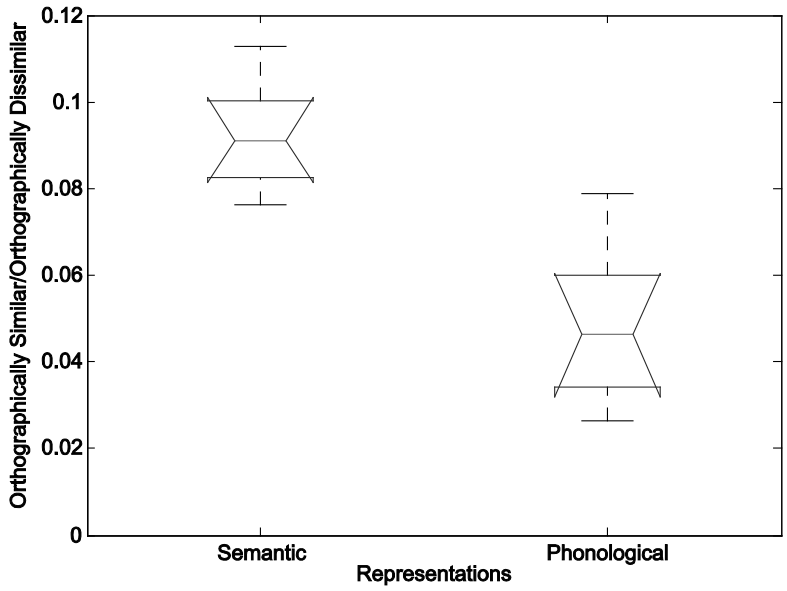

Figure 26: Difference as a consequence of literacy training in the ratio (VisEffect) of the distance between semantic and phonological layer activity when processing words that are orthographically similar compared to the distance between activation when processing words that are orthographically dissimilar during reading trials. [Semantic = distance between semantic layer activity; Phonological = distance between phonological layer activity] [2].

During Retention Tasks

Figure 27 displays the *VisEffect* ratio derived from phonological layer activity during phonological retention trials and semantic layer activity during semantic retention trials. To test whether orthographic similarity affected phonological and semantic representations in literate networks in the absence of activation of orthographic representations we tested whether the phonologically derived and semantically derived *VisEffect* ratios differed from zero using two one-sample t-tests. This analysis indicates that in the absence of activation of orthographic representations orthographic similarity does not influence the structure of semantic or phonological representations.

This analysis of the effects of orthographic similarity on phonological and semantic representations demonstrates that semantic and phonological representations of words that are orthographically similar become more distinct when their orthographic form is active, yet are not affected in the absence of activation of their orthographic form. To our knowledge, there are as yet no such studies that have investigated this issue.

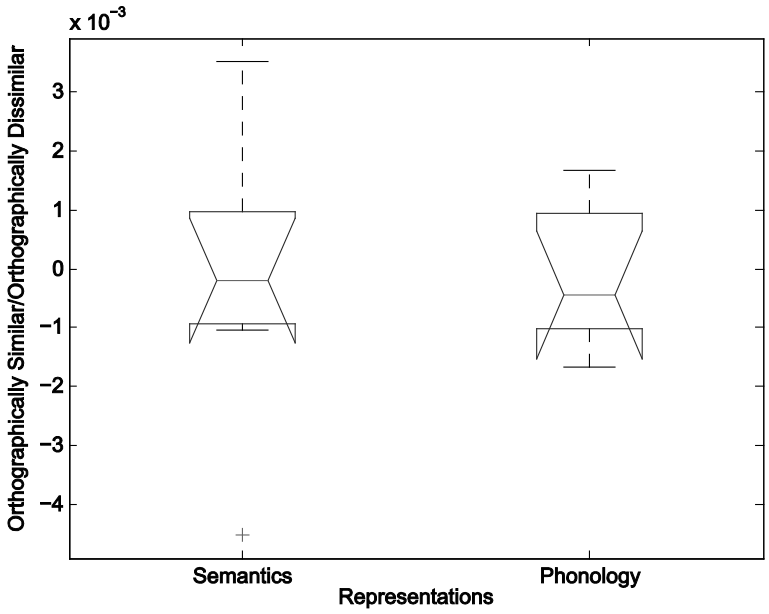

Figure 27: Difference as a consequence of literacy training in the ratio (VisEffect) of the distance between semantic and phonological layer activity when processing words that are orthographically similar compared to the distance between activation when processing words that are orthographically dissimilar during retention trials. [Semantic = distance between semantic layer activity; Phonological = distance between phonological layer activity] [2].

4. General Discussion

This study examined the scope of the triangle model of reading as a framework able to support reading in each of the world's major orthographic systems. Below we summarise the empirical findings such a universal model of reading is able to replicate while also discussing insights that can be derived from limitations of the current implementation.

The implementation outlined above provides an explicit description of how contrasts in processing can emerge as a consequence of differences in the statistical structure of the learning environment imposed by alternative orthographic systems. Previous experimental and computational attempts to isolate the effects of orthographic transparency have been limited by linguistic factors such as differences in semantic or phonological structure, or socio-cultural factors such as language exposure, teaching methods or student motivation. By assuming that reading across orthographic systems is supported by the same underlying architecture and statistical learning mechanisms we isolate the effects of orthographic transparency by manipulating the extent to which phonological or semantic structure encoded

within the orthography while holding phonological and semantic structure fixed. This approach allows us to demonstrate how orthographic transparency alone affects processing across development.

On the issue of acquisition our study replicates and offers explicit explanation for behavioural findings that show that phonological transparency aids decoding acquisition. Due to the componential phonological information encoded in the orthography networks trained on sub-syllabically transparent systems reached proficiency in phonological decoding prior to netowrks trained on logographic systems. The work further offers an empirically verifiable prediction of a positive effect of semantic transparency on decoding acquisition, with a marginal difference observed between networks trained on logographic opaque compared to logographic semantically transparent systems. Simulations also predict a positive influence of transparency on reading comprehension acquisition and faster comprehension acquisition rates in comparison to phonological decoding ability in opaque systems. However, the few studies that exist systematically examining the effects of transparency on reading comprehension suggest that transparency may reduce reading comprehension acquisistion rates (see Seidenberg, 2013 for review). Further studies are required to establish this relationship however should such a relationship exist our modelling constrains explanations to factors beyond the level of monosyllabic transparency or to issues relating to immaturity of phonological and semantic knowledge of a language prior to literacy training..

The current study also provides an explicit explanation of how processing differences can emerge as a consequence of orthographic transparency. We demonstrate how differences in activation of dorsal and ventral paths of the reading network both across development and in mature systems are emergent consequences of differences in orthographic transparency. Networks trained on orthographic systems that encoded sub-syllabic phonological structure displayed greater activation entering semantic processing regions via indirect paths (orthography to semantics via phonology) during word reading comprehension relative to activation entering via direct paths (orthography to semantics) compared to logographic and syllabic systems. By contrast networks trained on alphabetic and alphasyllabic systems displayed a greater level of activation entering phonological processing regions via direct paths (orthography to phonology) relative to activation entering such regions via indirect paths (orthography to phonology via semantics) during phonological decoding compared to networks trained on consonantal, syllabic and logographic systems. Our simulations generated the empirically untested prediction that alphasyllabic systems should generate a

dorsal bias during reading comprehension in processing and also predicts that differences in orthographic transparency should lead to contrasts, at least at early stages of reading acquisition, in the rate of activation of phonological and semantic information.

Finally, the current work indicates that should such an interactive activation architecture support reading then effects of orthographic transparency should be observed on both phonological and semantic processing irrespective of whether orthographic information is active. Our simulations demonstrate that learning mappings between orthography and phonology affects processing in phonological processing networks with effects on phonological representations modulated both by the manner in which phonological information is encoded in the orthography and the presence or absence of orthographic activation during processing. Similarly, semantic processing was also shown to be modulated by learning mappings between orthography and semantics when sub-lexical semantic information is embedded within the orthography.

In a recent commentary Frost, (2012) argues that a universal theory of reading should isolate what is invariant in orthographic processing across systems this should entail being able to describe what characterises human writing systems and the cognitive system that supports them. This set of universals should be small, general and abstract in order to fit all writing systems. This study examines the viability of the triangle model of reading as such a universal model and moves us closer towards isolating how orthographic structure may influence the reading processing providing us with a baseline as to the initial biases a given system brings. By building in further language specific features to the model we can explore how each feature specifically affects processing, that without this initial investigation at this level of abstraction it would not have been possible to isolate. Having outlined the successes of the current implementation we will next examine what can be learnt from its limitations.

The orthographic system that proved most difficult to implement given the limitations of our computational approach was the consonantal system. Due to the under-specification of the orthography consonantal systems generate a larger number of homographs. Without a top-down influence of semantic context the model was unable to distinguish between such items. This is clearly not an issue for the theoretical framework however it does place limitations on the validity of our results for consonantal systems as the increased value of pre-activated semantic information in consonantal systems is likely to have significant implications for the processing dynamics within the reading system.

A second orthographic system that generated a number of predictions that do not align with what is known in the literature was the syllabic system. For example we know that an individual reading Japanese in Kana (syllabic) displays increased activation of the dorsal path compared to when they are processing Kanji (logographic). Although we observed a modulation of dorsal vs ventral processing as a function of transparency this was not observed for syllabic vs logographic systems. Further, Asfaha et al, (2009) provides evidence to suggest that some syllabic systems may lead to faster decoding abilities in alphabetic systems, however our modelling results show similar rates of acquisition for both logographic and syllabic systems. We believe there are two factors that may result in these contrasting results. As the current implementation only models monosyllabic word reading all that defines differences between syllabic and logographic systems in the current implementation is a set of 25 homophones. There is therefore little difference between the complexity of learning an orthographic representation for every syllable in the language, opposed to every word in the language. The syllabic structure of Japanese however consists of approximately 100 distinct phonological units for each of which in a syllabic system there will be a distinct orthographic representation. Therefore, there is a significant decoding acquisition advantage in learning a transparent orthographic form of Japanese as only 100 distinct units are required to be learnt in order to decode all words in the language. Given this increased level of transparency we would predict that should this system be implemented in a multisyllabic version of the current model, it is likely to display behaviour closer to that of alphabetic and alphasyllabic systems than logographic systems both in terms of literacy acquisition rate and the division of labour, bringing behaviour closer to empirical findings. However, as transparency would still be weaker than alphabetic or alphasyllabic systems it is unlikely we would observe a decoding advantage for syllabic systems over alphabetic systems as has been observed (Asfaha et al., 2009).

A second factor that may influence predictions for syllabic systems are assumptions that the current model shares with many existing computational models of reading (e.g. Harm & Seidenberg, 1999, 2004; Coltheart, Rastle, Perry, Langdon & Ziegler 2001; Houghton & Zorzi, 2003) regarding the structure and acquisition of phonological representations. Within such models phonological information extracted from the speech signal is represented as possessing fine phonetic detail in which phonemic boundaries are clearly defined and variation of phonemes within types is minimal and independent of context. We know however that the speech signal is noisy and endemic with features such as co-articulation,

elision and reductions. Such features of the input are likely to have profound consequences for the emergent structure of phonological representations that are not currently captured within existing models of reading. A large body of empirical data from phonological awareness studies (see Morais & Kolinsky, 2001) indicates that literacy in alphabetic systems significantly alter at least explicit awareness of sub-syllabic phonological structure within the speech signal. Capturing an accurate depiction of the structure of emergent phonological representations is likely to greatly enhance the accuracy of predictions regarding the influence of transparency on reading acquisition and the impact of literacy acquisition on phonological processing. We do not however observe a graded effect of orthographic transparency on the granularity of phonological processing a property of the model that does not align with empirical data and theoretical models within the literature (see section 1). The current model also fails to capture a decoding advantage for syllabic systems over alphabetic systems whereas some empirical findings demonstrate this advantage (Asfaha et al., 2009). Asfaha et al., (2009) motivate their findings based on data that shows syllables are psychologically more accessible (Antony & Francis, 2005), a feature that is not captured in the current model. Both rates of acquisition and effects of literacy on phonological processing predicted by the model are likely to alter significantly should emergent phonological representations not contain the fine grained sub-phonemic detail that we (and others) have assumed to exist prior to literacy training.

Our investigation explores the scope of the triangle model of reading as a universal framework for supporting reading and capturing the effects of orthographic transparency on reading acquisition and processing more broadly. As authors we accept that each orthographic system is a product of the combined evolution a languages semantic, phonological and orthographic structure (Seidenberg, 2013; Frost 2012). However, we believe if we are to isolate the effects of transparency then computational modelling at such a level of abstraction provides the only means to do so. This study demonstrates that even without full implementation of phonological and semantic structure of languages a system that implements the same domain general mechanisms on distributed representations of phonological, semantic and orthographic information (Yang et al., 2009) is able to support reading across all of the world's major orthographic systems and further provides an explicit description for how a broad range of properties common to individual categories of orthographic system (e.g. faster phonological decoding acquisition and increased dorsal path processing bias as a consequence of increased phonological transparency) emerge from

constraints placed on the statistics of the learning environment by properties common to the orthographic system. As Frost (2012) argues "if the model indeed picks up the statistical regularities of the language and the expected reading behaviour emerges it most probably reflects the actual learning procedures of readers." Here we have implemented a level of abstraction that allows us to isolate the factors driving distinctions in behaviour without linguistic or socio-cultural confounds that have plagued previous studies.

References

Adrián, J. A., Alegria, J., & Morais, J. (1995). Metaphonological abilities of Spanish illiterate adults. International Journal of Psychology, 30(3), 329–351.

Alcock, K. J., Ngorosho, D., Deus, C., & Jukes, M. C. H. (2010). We don't have language at our house: disentangling the relationship between phonological awareness, schooling, and literacy. *British Journal of Educational Psychology, 80(1),* 55-76.

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science, 14(5),* 255-259.

Asfaha, Y. M., Kurvers, J., & Kroon, S. (2009). Grain size in script and teaching: Literacy acquisition in Ge'ez and Latin. Applied Psycholinguistics, 30(04), 709-724.

Bolger, D. J., Perfetti, C. A., & Schneider, W. (2005). Cross-cultural effect on the brain revisited: Universal structures plus writing system variation. Human brain mapping, 25(1), 92-104.

Brennan, C., Cao, F., Pedroarena-Leal, N., McNorgan, C., & Booth, J. R. (2013). Reading acquisition reorganizes the phonological awareness network only in alphabetic writing systems. *Human brain mapping, 34(12),* 3354-3368.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6(02), 201-251.

Bruck, M., Genesee, F., & Caravolas, M. (1997). A cross-linguistic study of early literacy acquisition. Foundations of reading acquisition and dyslexia: Implications for early intervention, 145-162.

Cheung, H., & Chen, H. C. (2004). Early orthographic experience modifies both phonological awareness and on-line speech processing. Language and Cognitive Processes, 19(1), 1-28.

Cheung, H., & Ng, L. (2003). Chinese reading development in some major Chinese societies: An introduction. In: McBride-Chang, C. & Chen, H. (eds.), *Reading Development in Chinese Children*. (pp. 3-17). Westport, CT: Praeger Publishers.

Cohen-Mimran, R. (2009). The contribution of language skills to reading fluency: A comparison of two orthographies for Hebrew. *Journal of Child Language*, 36(03), 657-672.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud.Psychological review, 108(1), 204.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. Psychological review, 108(1), 204.

Comrie, B. (2013). Writing Systems. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/141, Accessed on 2014-05-07.)

Das, T., Padakannaya, P., Pugh, K. R., & Singh, N. C. (2011). Neuroimaging reveals dual routes to reading in simultaneous proficient readers of two orthographies. Neuroimage, 54(2), 1476-1487.

De Gelder, B. D., & Vroomen, J. (1992). Auditory and visual speech perception in alphabetic and non-alphabetic Chinese-Dutch bilinguals. Advances in psychology, 83, 413-426.

de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology*, *91*(3), 450-476.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J. & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science, 330(6009),* 1359-1364.

Dehaene-Lambertz, G., Montavont, A., Jobert, A., Allirol, L., Dubois, J., Hertz-Pannier, L., & Dehaene, S. (2010). Language or music, mother or Mozart? Structural and environmental influences on infants' language networks. Brain and Language, 114(2), 53–65. doi:10.1016/j.bandl.2009.09.003

Devauchelle, A. D., Oppenheim, C., Rizzi, L., Dehaene, S., & Pallier, C. (2008). Sentence syntax and content in the human temporal lobe: an fMRI adaptation study in auditory and visual modalities. Journal of Cognitive Neuroscience, 21(5), 1000-1012.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010). Are there mental lexicons? The role of semantics in lexical decision. *Brain research*, *1365*, 66-81.

Durgunoglu, A. Y. (2006). How the language's characteristics influence Turkish literacy development. Handbook of orthography and literacy, 219-230.

Frith, U., Wimmer, H., & Landerl, K. (1998). Differences in phonological recoding in German-and English-speaking children. Scientific Studies of Reading, 2(1), 31-54.

Frost, R. (2012). A universal approach to modeling visual word recognition and reading: Not only possible, but also inevitable. Behavioral and Brain Sciences, 35(05), 310-329.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. Journal of Experimental Psychology: Human Perception and Performance, 13(1), 104.

Goswami, U., Gombert, J. E., & de Barrera, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French, and Spanish. Applied Psycholinguistics, 19(01), 19-52.

Hanley, R., Masterson, J., Spencer, L., & Evans, D. (2004). How long do the advantages of learning to read a transparent orthography last? An investigation of the reading skills and reading impairment of Welsh children at 10 years of age. The Quarterly Journal of Experimental Psychology: Section A, 57(8), 1393-1410.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review, 106(3),* 491-528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662-720.

Hirshorn, E. A., & Fiez, J. A. (2014). Using artificial orthographies for studying cross-linguistic differences in the cognitive and neural profiles of reading. Journal of Neurolinguistics, 31, 69-85.

Hoonhorst, I., Medina, V., Colin, C., Markessis, E., Radeau, M., Deltenre, P., & Serniclaes, W. (2011). Categorical perception of voicing, colors and facial expressions: A developmental study. *Speech communication, 53(3),* 417-430.

Houghton, G., & Zorzi, M. (2003). Normal and impaired spelling in a connectionist dual-route architecture. Cognitive Neuropsychology, 20(2), 115-162.

Hudson, R. F., Torgesen, J. K., Lane, H. B., & Turner, S. J. (2012). Relations among reading skills and sub-skills and text-level reading proficiency in developing readers. *Reading and Writing*, 25(2), 483-507.

Huettig, F., & Mishra, R. K. (2014). How literacy acquisition affects the illiterate mind - A critical examination of theories and evidence. *Language and Linguistics Compass, 8*(10), 401-427.

Huettig, F., Singh, N., & Mishra, R. K. (2011). Language-mediated visual orienting behavior in low and high literates. *Frontiers in Psychology, 2,* 285.

Hulme, C., Snowling, M., Caravolas, M., & Carroll, J. (2005). Phonological skills are (probably) one cause of success in learning to read: A comment on Castles and Coltheart. *Scientific Studies of Reading, 9(4),* 351-365.

Jobard, G., Crivello, F., Tzourio-Mazoyer, N., 2003. Evaluation of the dual route theory of reading: a metanalysis of 35 neuroimaging studies. *Neuroimage, 20*, 693–712.

Katz, L., & Feldman, L. B. (1981). Linguistic coding in word recognition: Comparisons between a deep and a shallow orthography. In A. M. Lesgold & C. A. Perfetti (Eds.), Interactive Processes in Reading. Hillsdale, NJ: Erbaum, 85-106.

Kidd, J. C., Shum, K. K. M., Ho, C. S. H., & Au, T. K. F. (2014). Phonological Representations and Early Literacy in Chinese. Scientific Studies of Reading, (ahead-of-print), 1-25.

Kiyosawa, M., Itoh, M., Nakagawa, Y., Kobayashi, N., Tamai, M., 1995. Effect of kanji and kana reading on cerebral blood flow patterns measured by PET. Jpn. J. Ophthalmol. 39, 198–205.

Krakow, R. A. (1999). Physiological organization of syllables: a review. *Journal of Phonetics*, 27(1), 23-54.

Levy, J., Pernet, C., Treserras, S., Boulanouar, K., Berry, I., Aubry, F., Demonet, J.F., Celsis, P., 2008. Piecemeal recruitment of left-lateralized brain areas during reading: a spatio-functional account. *Neuroimage, 43*, 581–591.

Levy, J., Pernet, C., Treserras, S., Boulanouar, K., Aubry, F., Demonet, J.F., Celsis, P., 2009. Testing for the dual-route cascade reading model in the brain: an fMRI effective connectivity account of an efficient reading style. *PLoS One, 4*, e6675.

Loureiro, C. D. S., Willadino Braga, L., Souza, L. D. N., Queiroz, E., & Dellatolas, G. (2004). Degree of illiteracy and phonological and metaphonological skills in unschooled adults. *Brain and Language, 89(3),* 499-502.

Monzalvo, K., & Dehaene-Lambertz, G. (2013). How reading acquisition changes children's spoken language network. *Brain and Language, 127*(3), 356-365.

Monaghan, P. & Ellis, A.W. (2010). Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language, 63,* 506-525.

Monaghan, P., Shillcock, R., & McDonald, S. (2004). Hemispheric asymmetries in the split-fovea model of semantic processing. *Brain and Language*, *88*(3), 339-354.

Morais, J., & Kolinsky, R. (2001). The literate mind and the universal human mind. In Dupoux, E., & Mehler, J. (Eds). Language, brain and cognitive development: Essays in Honor of Jacques Mehle. MIT, Cambridge, Mass, 463-480.

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition, 7(4),* 323-331.

Nag, S. (2007). Early reading in Kannada: The pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading, 30*(1), 7-22.

Nakamura, K., Kuo, W. J., Pegado, F., Cohen, L., Tzeng, O. J., & Dehaene, S. (2012). Universal brain systems for recognizing word shapes and handwriting gestures during reading. *Proceedings of the National Academy of Sciences, 109*(50), 20762-20767.

Nation, K., & Cocksey, J. (2009). The relationship between knowing a word and reading it aloud in children's word reading development. *Journal of Experimental Child Psychology, 103*(3), 296-308.

Pattamadilok, C., Knierim, I. N., Duncan, K. J. K., & Devlin, J. T. (2010). How does learning to read affect speech perception? *Journal of Neuroscience, 30(25),* 8435-8444.

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S.F., Cotelli, M., Cossu, G., Corte, F., Lorusso, M., Pesenti, S., Gallagher, A., Perani, D., Price, C., Frith, C.D., Frith, U., 2000. A cultural effect on brain function. *Nature Neuroscience, 3*, 91–96.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, 1(2), 263–269.

Perfetti, C. A., Liu, Y., & Tan, L. H. (2005). The lexical constituency model: some implications of research on Chinese for general theories of reading. *Psychological Review, 112*(1), 43-59.

Perre, L., Pattamadilok, C., Montant, M., & Ziegler, J. C. (2009). Orthographic effects in spoken language: On-line activation or phonological restructuring?. *Brain Research, 1275,* 73-80.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114*, 273-315.

Perry, C., Ziegler, J. C., & Zorzi (2010). Beyond single syllables: Large-scale modelling of reading aloud with the connectionist dual process (CDP++) model. *Cognitive Psychology*, 61, 2, 106-151.

Perry, C., Ziegler, J. C., & Zorzi, M. (2014a). CDP++.Italian: Modelling sublexical and supralexical inconsistency in a shallow orthography. *Plos One, 9(4),* e94291.

Perry, C., Ziegler, J. C., & Zorzi, M (2014b). When silent letters say more than a thousand words: An implementation and evaluation of CDP++ in French. *Journal of Memory and Language, 72*, 98-115.

Petersson, K. M., Ingvar, M., & Reis, A. (2009). Language and literacy from a cognitive neuroscience perspective. In D. Olsen, & N. Torrance (Eds.), *Cambridge handbook of literacy* (pp. 152-181). Cambridge: Cambridge University Press.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56-115.

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*, 62-88.

Price, C. J., & Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, *15*(6), 246-253.

Pugh, K.R., Mencl, W.E., Jenner, A.R., Katz, L., Frost, S.J., Lee, J.R., Shaywitz, S.E., & Shaywitz, B.A. (2000). Functional neuroimaging studies of reading and reading disability (developmental dyslexia). Ment. Retard. Dev. Disabil. Res. Rev. 6, 207–213.

Pugh, K.R.,Mencl,W.E., Jenner, A.R., Katz, L., Frost, S.J., Lee, J.R., Shaywitz, S.E., & Shaywitz, B.A. (2001). Neurobiological studies of reading and reading disability. J. Commun. Disord. 34, 479–492.

R Development Core Team. (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Read, C., Yun-Fei, Z., Hong-Yin, N., & Bao-Qing, D. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition*, 24(1), 31-44.

Reis, A., & Castro-Caldas, A. (1997). Illiteracy: A cause for biased cognitive development. *Journal of the International Neuropsychological Society, 3(05),* 444-450.

Richardson, F.M., Seghier, M.L., Leff, A.P., Thomas, M.S., & Price, C.J., (2011). Multiple routes from occipital to temporal cortices during reading. *Journal of Neuroscience, 31*, 8239–8247.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533-536.

Scliar-Cabral, L., Morais, J., Nepomuceno, L. & Kolinsky, R. (1997). The awareness of phonemes: So close-so far away. *International Journal of Psycholinguistics, 13,* 211-240.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. Psychological review, 96(4), 523.

Seidenberg, M. S. (2013). The science of reading and its educational implications. *Language Learning and Development, 9*, 331-360.

Serniclaes, W., Ventura, P., Morais, J., & Kolinsky, R. (2005). Categorical perception of speech sounds in illiterate adults. *Cognition, 98*, B35–B44.

Seymour, P. H., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143-174.

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Mencl, W. E., Fulbright, R. K., Skudlarski, P., ... & Gore, J. C. (2002). Disruption of posterior brain systems for reading in children with developmental dyslexia. Biological psychiatry, 52(2), 101-110.

Shu, H., Peng, H., & McBride-Chang, C. (2008). Phonological awareness in young Chinese children. *Developmental Science, 11(1),* 171-181.

Smith, A., Monaghan, P., & Huettig, F. (2014a). Literacy effects on language and vision: Emergent effects from an amodal shared resource (ASR) computational model. Cognitive Psychology, 75, 28-54. doi:10.1016/j.cogpsych.2014.07.002.

Smith, A., Monaghan, P., & Huettig, F. (2014b). Examining strains and symptoms of the 'Literacy Virus': The effects of orthographic transparency on phonological processing in a connectionist model of reading. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci 2014). Austin, TX: Cognitive Science Society.

Snowling, M. J., & Hulme, C. (Eds.). (2005). *The science of reading: A handbook*. Oxford, UK: Blackwell.

Tan, L. H., Laird, A. R., Li, K., & Fox, P. T. (2005). Neuroanatomical correlates of phonological processing of Chinese characters and alphabetic words: A meta-analysis. *Human brain mapping*, 25(1), 83-91.

Tong, X., & McBride-Chang, C. (2010). Chinese-English biscriptal reading: Cognitive component skills across orthographies. *Reading and Writing*, 23(3-4), 293-310.

Tong, X., McBride-Chang, C., Shu, H., & Wong, A. M. (2009). Morphological awareness, orthographic knowledge, and spelling errors: Keys to understanding early Chinese literacy acquisition. *Scientific Studies of Reading*, 13(5), 426-452.

Treiman, R., & Zukowski, A. (1991). Levels of phonological awareness. In S. A. Brady, & D. P. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman,* (pp. 67-83). Oxford: Routledge.

Ueno, T., & Lambon Ralph, M. A. (2013). The roles of the "ventral" semantic and "dorsal" pathways in conduite d'approche: A neuroanatomically-constrained computational modeling investigation. *Frontiers in Human Neuroscience*, *7,* 422.

Winskel, H., & Iemwanthong, K. (2010). Reading and spelling acquisition in Thai children. *Reading and Writing, 23*(9), 1021-1053.

Yang, J., Shu, H., McCandliss, B. D. & Zevin, J. D. (2013). Orthographic influences on division of labor in learning to read Chinese and English: Insights from computational modeling. *Bilingualism: Language and Cognition, 16(2),* 354-366.

Yang, J., McCandliss, B. D., Shu, H., & Zevin, J. D. (2009). Simulating language-specific and language-general effects in a statistical learning model of Chinese reading. *Journal of Memory and Language, 61(2),* 238-257.

Yang, J., Zevin, J. D., Shu, H., McCandliss, B. D., & Li, P. (2006). A "triangle model" of Chinese reading. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Zhou, Y. G. [周有光] (1978). To what extent are the "phonetics" of present-day Chinese characters still phonetic? [现代汉字声旁的表音功能问题] *Zhongguo Yuwen [中国语文], 146*, 172-177.

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading a cross-language investigation. *Psychological Science*. 21(4), 551-559.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin, 131(1),* 3-29.

Appendix

Neural networks simulations were performed using Mikenet version 8.0 developed by M. W. Harm (www.cnbc.cmu.edu/~mharm/research/tools/mikenet/), a collection of libraries written in the C programming language for implementing and training connectionist networks.

Networks were trained using the continuous recurrent backpropagation through time training algorithm provided in Mikenet (crbp.c) which implements Pearlmutter (1989). Unit activation was calculated using a logistic activation function and following the implementation of Harm and Seideberg, (2004) error was calculated using the cross-entropy measure and an error radius of 0.1. Also replicating the implementations reported in Harm and Seidenberg, 2004 time averaged input networks were used. This implements the assumption that greater activation generates faster responses and that the network is pressured to produce the correct response rapidly. Time within the network was modelled using 12 samples and an integration constant of 0.25. All other parameters were set to the default values implemented in Mikenet version 8.0.

Mixed effects model analysis was performed using the R (version 3.1.0; R Development Core Team, 2009) libraries lme4 (version 1.1-6) and languageR (version 1.4.1).

Chapter 7

Summary & Conclusions

Within this chapter we summarise the main findings of each chapter in this thesis and place these combined results in a broader context by exploring conclusions that can be drawn from this collective set of studies. We also discuss potential future avenues of investigation that the products of these studies invite.

Summary of results

This thesis explored the implications of interactive models of language processing that aim to describe the multimodal nature of spoken and written word processing. I used emergent neural network models, derived from the parallel distributed processing tradition (Rumelhart, McClelland & the PDP Research Group, 1986; see Rogers & McClelland, 2014; McClelland et al., 2014 for review) that captured the interaction of visual, semantic and auditory information processing streams. Studies examined how the structure of information within the multimodal learning environment may shape processing and behaviour over both the course of development and in mature systems. These properties of the models were then evaluated for their adequacy and explanatory scope against existing behavioural and neural data sets in addition to novel data from visual world studies conducted as part of this thesis.

Chapters 2 and 3 detail the motivations and assumptions that support the multimodal integration model of language mediated visual attention (MIM). MIM offers an explicit description of the interaction between visual and linguistic processing streams that is probed by studies of language mediated visual attention. It describes explicitly the connection between the input (visual and auditory stimuli) and the output (eye gaze) of the system that supports this behaviour. The model uses a hub-and-spoke architecture, derived from models of semantic processing (Rogers et al. 2004; Plaut, 2002; Lambon-Ralph & Patterson, 2008;

Dilkina, McClelland & Plaut, 2008), to allow visual, semantic and auditory information to interact in parallel via a central connecting resource. Minimal additional constraints are placed on the flow of information within the system, therefore emergent processing and behavioural characteristics of the model are dependent on constraints imposed by interaction with the multimodal learning environment. The structural properties of the environment that affect behaviour or processing within the model are also defined explicitly. Within the model these properties relate to basic structural properties of the information available or task demands, motivated by the developmental literature (McMurray, Horst, & Samuelson, 2012; Roy, 2005; Smith & Yu, 2008; Vouloumanos & Werker, 2009; Yu, Ballard, & Aslin, 2005), that allow such a model to learn simple cross modal mappings.

In chapter 2 I demonstrate that MIM is able to replicate a broad range of distinct visual, semantic and phonological similarity effects on language mediated visual attention (see table 1, rows 1-6). The model displayed increased fixation of items in the visual display that overlapped with a spoken target word in either a visual, semantic or phonological dimension compared to items that were unrelated to the target word. This occurred both when the visual object relating to the target word was present and absent from the visual display. Critically, the model displayed modality specific characteristics of these effects. For example, phonological onset competitors were fixated more at early stages of processing of the spoken word (Allopenna et al., 1998; Huettig & McQueen, 2007). By contrast, phonological rhyme effects were shown to be weaker and later and to be dependent on the presence of noise within the speech signal encountered within the learning environment (Allopenna et al., 1998; McQueen & Huettig, 2012; McQueen & Viebahn, 2007). Demonstrating the models ability to capture these results provided a necessary precursor before extending to model multiple interactive effects as examined in chapters 3-5.

Chapter 3 examined the model's ability to replicate the time course of multiple interactive effects of language mediated visual attention both over the course of processing a single spoken word and over the course of development (Table 1, rows 7-10). Further, investigations detailed in chapter 3 probed the internal processing of the model both over the course of development and in the mature system in order to offer explanation for observed effects and inform inferences regarding properties of the representations and cognitive architecture that support multimodal cognition.

Analysis of representations developed within the integrative hub of MIM showed that at early stages of development there was evidence for amodal processing. However, with increased training networks incorporated an increasing number of aspects of the input modality into the representation, such that at the end of training there was no evidence for amodal processing within the integrative layer. This analysis extends previous modelling investigations that utilise a hub-and-spoke architecture (Rogers et al. 2004; Plaut, 2002; Lambon-Ralph & Patterson, 2008; Dilkina, McClelland & Plaut, 2008) showing that amodal representations do not necessarily result from such an architecture. Further, this data is also consistent with studies of embodied cognition that argue that properties of the representations activated are intimately related to the input and output modalities involved (Barsalou, 1999; Wilson, 2002; see Barsalou, 2008 for review).

Results reported in chapter 3 demonstrate that the emergent behavioural properties of MIM replicate the complex time course dynamics of multimodal interactive effects on language mediated visual attention both over the course of single word processing and over the course of development (Mani & Huettig, submitted; Huettig & McQueen, 2007). For example, the model displayed a progressively greater semantic effect at later stages of word processing and a progressively earlier distribution of phonological effects over the course of development. These results therefore argue that such observed properties of language mediated visual attention are likely to reflect an increase in the strength of cross-modal associations that result from increased language exposure. In addition, distinct visual, semantic and phonological effects observed in multi-competitor scenes (e.g. Huettig & McQueen, 2007) were shown to involve the interaction of information activated across all three modalities represented in the model. Therefore, they could not be described in terms of mapping at a single level of representation as previous descriptive models had proposed (Altmann & Kamide, 2007; Altmann & Mirkovic, 2009; Huettig, Olivers & Hartsuiker, 2011; Huettig, Olivers, & Mishra, 2012). Thus, due to the parsimony of the model it was possible to isolate such complex behavioural properties as emergent consequences of a shared resource responding to constraints imposed by the structure of the learning environment.

Results reported in chapter 4 extend compatibility of MIM beyond replication of existing word level effects to generate hypotheses and predict behavioural patterns in novel data sets. Within this study I investigated the influence of information carried in the phonological rhyme of words during spoken word processing under conditions in which visual and semantic information is also available to constrain processing. MIM was used to simulate

visual world conditions in which participants are exposed to scenes containing distinct visual, semantic and phonological rhyme competitors. MIM, which implements an architecture in which visual, semantic and phonological information is integrated in parallel, predicted earlier and stronger visual and semantic competitor effects compared to phonological rhyme effects. These predictions were confirmed in two visual world experiments that showed that visual and semantic information was activated rapidly to constrain processing such that similarity in the phonological rhyme of words had no observable influence on behaviour even when phonological rhyme competitors only differed from target words in their initial phoneme.

After demonstrating MIM's adequacy as a model of language mediated visual attention in chapters 2 – 4, in chapter 5 I applied the model to investigate individual differences in language processing. Specifically I examined whether recent distinctions observed between high and low literate populations in sensitivity to semantic and phonological competitors in language mediated visual attention (Huettig, Singh & Mishra, 2011) was compatible with theoretical perspectives that argue that literacy leads to finer grain phonological processing (Muneaux & Ziegler, 2004; Taft & Hambly, 1985; Taft, 2006; Ziegler & Goswami, 2005). Simulations within MIM show that changes to the granularity of phonological representations within the model modulates the nature of the phonological competitor effect observed. Further by implementing coarse grain phonological representations in MIM the model is able to generate fixation behaviour similar to that displayed by low literates. Thus, supporting claims that literacy leads to finer grain phonological processing during online speech processing.

Finally, in chapter 6 I extend investigations of the effects of literacy acquisition on language processing by exploring within a computational model of reading the emergent effects of orthographic transparency on the reading system and wider language processing system. Specifically, I trained a connectionist implementation of the triangle model of reading on a range of orthographic systems that represented the range of the world's writing systems while controlling for phonological and semantic structure.

Results detailed in chapter 6 demonstrate the viability of the triangle model of reading as a universal model of reading (see Frost, 2012 for review) able to support reading across the breadth of the world's orthographic systems. By analysing for each system the models ability to map from orthography to phonology and from orthography to semantics over the course of

training I showed a graded effect of phonological and semantic transparency on both the acquisition of phonological decoding abilities and reading comprehension abilities. This data therefore offers support for arguments that transparency aids phonological decoding acquisition (Goswami, Gombert & De Barrera, 1998; Seymour, Aro & Erskine, 2003; Bruck, Genesee & Caravolas, 1997; Hanley, Masterson, Spencer & Evans, 2004) and due an absence of empirical data within the literature generates the prediction that transparency also aids comprehension acquisition (see Seidenberg, 2013 for review).

Further, analysis of the flow of activation within the model also showed a graded effect of orthographic transparency on the distribution of activation across indirect and direct paths during both phonological decoding and reading comprehension. Capturing this effect within the model allows us to provide an explicit description of how differences observed in activation levels of ventral (direct) and dorsal (indirect) paths within the reading system observed in neuroimaging studies (Paulesu et al., 2000; Kiyosawa et al., 1995) can emerge as a consequence of the same underlying architecture and learning mechanism configuring around orthographic systems that differ in their semantic and phonological transparency.

Finally, analysis of the structure of phonological and semantic processing with the model after extended exposure to literacy training demonstrated an effect of both phonological and semantic transparency on processing in both the presence and absence of orthographic activation. This is consistent with reports that phonological processing regions become restructured as a result of literacy training (Dehaene et al., 2010; Perre et al., 2009; Pattamadilok et al., 2010) and generates the predictions that not only should this extend to the restructuring of semantic processing regions engaged in the reading process but further that the nature of such effects should be modulated by the level to which phonological and semantic structure is encoded within the orthography.

Critically, as orthographic structure was the only factor to differ between simulations it was possible, within this study, to isolate the effects of orthographic structure on the reading system, something that has not been possible in previous behavioural and neural imaging studies due to co-varying linguistic, socio-economic and socio-cultural factors.

Language mediated visual attention

Studies detailed in chapters 2-5 of this dissertation demonstrate that the assumptions implemented in MIM are sufficient to account for a broad range of word level effects

reported in the language mediated visual attention literature (see table 1). The model describes how such features of behaviour are emergent properties both of basic structural features of the representations and computational properties of the mappings performed between them, thus arguing that many word level features reported in the literature are necessary consequences of developing multimodal knowledge of items.

Table 1: Properties of language mediated visual attention replicated by the Multimodal Integration Model. (Onset = phonological onset competitor, Rhyme = phonological rhyme competitor, Visual = visual competitor, Semantic = semantic competitor, Unrelated = unrelated distractor, Target = target object).

Study	Scene				Population
Authors (year)	Item 1	Item 2	Item 3	Item 4	
Allopenna et al. (1998)	Target	**Onset**	**Rhyme**	Unrelated	Literate adults
Dahan & Tanenhaus (2005)	Target	**Visual**	Unrelated	Unrelated	Literate adults
Huettig & Altmann (2007)	**Visual**	Unrelated	Unrelated	Unrelated	Literate adults
Yee & Sedivy (2006)	Target	**Semantic**	Unrelated	Unrelated	Literate adults
Huettig & Altmann (2005)	**Semantic**	Unrelated	Unrelated	Unrelated	Literate adults
Mirman & Magnuson (2009)	Target	**Near Semantic**	**Far Semantic**	Unrelated	Literate adults
Huettig & McQueen (2007)	**Onset**	**Semantic**	**Visual**	Unrelated	Literate adults
Smith, et al. (submitted)	**Rhyme**	**Semantic**	**Visual**	Unrelated	Literate adults
Huettig, Singh & Mishra (2011)	**Semantic**	**Phonological**	Unrelated	Unrelated	Low-literate adults
Mani & Huettig (submitted)	**Semantic**	**Phonological**	Unrelated	Unrelated	Children (age = 2, 4, 6, 8)

MIM fills an explanatory gap identified by many within the literature (Ferriera & Tanenhaus, 2007; Huettig, Olivers & Hartsuiker, 2011; Anderson et al., 2011) offering an explicit description of the role of language, memory, vision and attention in connecting processing of concurrent visual and auditory signals to variation in eye gaze behaviour.

In offering such a description the model develops our understanding of the connection between variation in gaze in visual world studies and the extent to which this reflects processing in the underlying system, thus reducing ambiguity in the linking hypothesis. Previous theoretical models of word level effects observed in the visual world paradigm had framed explanations in terms of mapping at discrete levels of representation (visual: Dahan & Tanehaus, 2005; phonological: Tanenhaus, Magnuson, Dahan & Chambers, 2000; visual, semantic and phonological: Huettig & McQueen, 2007) and dependant on modular systems and/or cascading information between modalities (Altmann & Kamide, 2007; Altmann & Mirkovic, 2009; Huettig, Olivers & Hartsuiker, 2011; Huettig, Olivers, & Mishra, 2012). However, no model of this type has provided an explicit description of how such a system is capable of generating such patterns of effects. MIM by contrast offers an explicit

implementation of an alternative perspective and has demonstrated its adequacy in offering explanation for a broad range of word level effects. Within this parallel interactive system effects result from multiple information sources taking different times to co-activate one another, as a consequence not of architectural constraints but due to constraints from the representations themselves. The model allows gaze to be influenced by the parallel integration of information at all levels of representation and across all modalities, yet it is still influenced by temporal properties of the flow of information through the system and the interaction of information across modalities. MIM thus reframes the debate focussing on the interaction of information across modalities rather than activation of information within distinct modalities. Within this new framework it makes little sense to interpret effects as a consequence of mapping at a single level of representation instead to explain any effect it is necessary to describe the multimodal system as a whole.

Studies of the models adequacy do however identify areas of the model that may benefit from additional development that may otherwise limit the scope of further investigations. For example, the current implementation does not allow changes in gaze to alter the nature of the visual input, although clearly this is a property of the visual system. We know that from approximately 6 months of age infants can control saccades and fixation behaviour for example to smoothly track a moving object in the local visual environment (see Trueswell, 2008 for review). Such properties of the system are likely to minimally distort the developmental profile displayed by the model. For example, the rate of acquisition of vision to semantic mappings is likely to rapidly increase should the model be able to focus on a single item. Changing this property of the model is likely to impact on the manner in which information is extracted from a visual display which may also benefit from further development. The model currently represents the visual signal in full acuity from signal onset. This ignores any temporal dimension to the nature of the information extracted from the signal for example whether there is a graded availability of low and then high spatial frequency visual information (Bar, 2007).

Further, development of the representation of the speech signal within the model may also benefit future investigations. The current implementation underestimates the complexity of the speech signal as it does not implement multi-word inputs, co-articulation across word boundaries or across phoneme boundaries. These properties of speech as we discuss later in this chapter are likely to impact predictions regarding emergent representations and processing within the system. These are all extensions however that can be easily

accommodated by the architecture outlined in this thesis and therefore this framework provides a means of assessing in isolation their influence on behaviour and processing.

Given the successes of MIM outlined in chapters 2-5 I believe the model to offer a meaningful proxy of the cognitive system supporting language mediated visual attention. Operating at a level of abstraction that allows researchers to maintain traction and understanding on a wide range of issues relating to language mediated visual attention and broader aspects of the cognitive system exposed by this behaviour. Initial investigations with this model, outlined in this thesis, focussed on understanding the emergent properties of basic structural features of the input. Having detailed this link and demonstrated the models adequacy MIM now provides a baseline framework in which hypotheses regarding the relationship between behaviour and specific properties of the architecture, representations or learning environment can be implemented and tested in a tractable manner.

Multimodal language processing

MIM is not only a model of language mediated eye gaze but also provides an explicit description of the system supporting multimodal language processing. It frames spoken word recognition and comprehension in terms of multimodal constraint satisfaction, proposing that concurrent phonological, semantic and visual information are integrated in parallel. Further, as a computational model it describes explicitly how such a system can be supported by a cognitively plausible architecture (Rogers & McClelland, 2014) and thus can be used to generate empirically verifiable predictions of the consequences of these assumptions.

Given the ambiguity present within natural language (Piantadosi, Tily & Gibson, 2012) we know that the language processing system must rapidly accommodate cues from the current multimodal evidential landscape in order for communication to progress efficiently and effectively. However, our understanding of the architecture that supports this process is weakly defined with models of speech recognition frequently overlooking this multimodal aspect of the speech recognition problem (e.g. Luce et al., 2000; McClelland & Elman, 1986; Norris & McQueen, 2008; Scharenborg & Boves, 2010).

As investigations detailed within this dissertation illustrate inferring properties of the underlying system from behavioural or neural data is fraught with danger due to the complexity of this dynamic multimodal system and the richness of the dynamic multimodal environment. For this reason explicit implementation of hypotheses in computational models

of the type utilised in this thesis are required in order to maintain traction. MIM aims to offer a baseline model in which further assumptions regarding the structure of the underlying system can be tested effectively. One example of an issue appropriate for further investigation within the framework provided by MIM is the influence of multimodal information on perception. The current parsimonious architecture implemented in MIM does not allow information to feedback to visual or auditory input levels. Implementing feedback to these levels would allow processing of the input, or perceptual processing to be immediately influenced by the current multimodal evidential landscape (Anderson et al., 2011). In this way it may be possible to generate empirically verifiable predictions regarding the effects of feedback on for example anticipatory eye movements that would in turn offer a means of testing the validity of such assumptions regarding the underlying system

Effects of literacy on language processing

As discussed in chapters 3, 5 and 6 the literature is divided as to the effects of literacy on online phonological processing. Although both simulations in MIM and implementations of the triangle model of reading detailed in chapter 6 predict effects of literacy on online phonological processing, the nature of the effects predicted varied significantly.

Results reported in chapter 3 show that training on fine grain phonological representations necessarily leads to increasingly earlier and concentrated time locked phonological onset effects similar to those displayed by child and adult populations exposed to literacy training (Mani & Huettig, submitted; Huettig, & McQueen, 2007). As it is difficult to justify illiterates being exposed to less co-occurrence of phonological and visual patterns, results in chapter 5 show that in order for MIM to display behaviour consistent with that of low-literates (i.e. they do not display time locked sensitivity to phonological relationships between the speech signal and visually presented objects) phonological representations must be coarse grain. Such predictions of literacy leading to finer grain phonological processing are consistent with theoretical models that argue that changes in the granularity of processing are driven by training on orthographic representations in which the fine grain phonological structure of the given language is encoded (Ziegler & Goswami, 2005). However, although effects on phonological processing were shown to be influenced by orthographic transparency in chapter 6 the nature of the effects was small in comparison the changes to representations implemented in chapter 5.

A potential explanation for this mismatch between model behaviour and empirical findings may be found in the assumptions underlying the representations of speech and pre-literate phonological processing within both MIM and the reading model implemented in chapter 6. Both models assume that pre-literate phonological representations are established through learning to retain in memory a speech signal in which phonemes are represented as discrete units with no pronunciation variation within phoneme categories and no effect of phonological context. These assumptions are also shared with many existing computational models of reading and speech processing (Harm & Seidenberg, 1999, 2004; Coltheart, Rastle, Perry, Langdon & Ziegler 2001; Houghton & Zorzi, 2003; Gaskell & Marslen-Wilson, 1997; Rogers et al., 2004; Dilkina, McClelland & Plaut, 2008). However, contrary to these assumptions the speech signal is inherently noisy, littered with features such as co-articulation, elision and reductions (e.g. Krakow, 1999; Browman & Goldstein, 1989; see Farnetani & Recasens, 1997 for review) that are likely to have profound implications for the emergent pre-literate and post-literate phonological representations such systems develop (see Plaut & Kello, 1999). Therefore, to further investigate the emergent structure of phonological representations and the impact of literacy training on these representations it is likely necessary to extend implementations to include such properties of the speech signal. Both MIM and the model of reading implemented in chapter 6 offer a framework in which the effects of these additional properties of the speech signal can be investigated.

A gap remains in our understanding of the extent to which literacy alters online speech processing (and broader aspects of cognitive processing) and the mechanism through which it exerts an influence. To date most influential models of speech processing do not accommodate a role for orthographic influences on processing (e.g., Cohort Model, Marslen-Wilson & Tyler, 1980; MERGE, Norris, McQueen, & Cutler, 2000; Shortlist B, Norris & McQueen, 2008; TRACE, McClelland & Elman, 1986), however through the extension of models such as those used within this thesis in the manner outlined above it may be possible to gain traction on such issues.

Accelerating progress with computational models

In the introduction to this thesis the following quote by Berkeley in 1705 was used to illustrate the major debates within cognitive science that have run for centuries and are engaged by the investigations detailed within this thesis.

“No sooner do we hear the words of a familiar language pronounced in our ears but the ideas corresponding thereto present themselves to our minds: in the very same instant the sound and the meaning enter the understanding: so closely are they united that it is not in our power to keep out the one except we exclude the other also. We even act in all aspects as if we heard the very thoughts themselves.” G. Berkeley, An Essay Towards a New Theory of Vision, Dublin, 1709.

The models of language processing described within this dissertation offer explicit implementations of distinct positions on each of these major debates. For example, at what stage of language processing can information in one modality influence processing in another? “…in the very same instant the sound and the meaning enter the understanding…”, our implementations demonstrate that an architecture in which visual, semantic and phonological information interact in parallel is able to generate behaviour consistent with behaviour displayed by a broad range of populations at various stages of development. “…so closely are they [sound and meaning] united that it is not in our power to keep out the one except we exclude the other also” probing the emergent representations within our model of language mediated visual attention demonstrated that behaviour consistent with empirical data could be generated by a system that was not dependent on amodal representations. “We even act in all aspects as if we heard the very thoughts themselves” models implemented within this thesis also demonstrate how language processing can be grounded in our perceptual experience of the shared world that surrounds us, and how variation between populations in the representations developed to support language processing can emerge through differences in the structure of the information to which they are exposed.

In order to move such long standing debates forward and distinguish between competing hypotheses regarding the structure of the underlying system it is crucial that we are able to generate accurate predictions of the consequences of specific hypotheses that can then be tested in empirical studies. Implementation in computational models, as this thesis demonstrates, offers a rare means of acquiring such accurate predictions of the influence of explicitly defined properties of such complex and dynamic systems as they interact with a rich and dynamic multimodal world.

However, for such an approach to be successful an appropriate level of abstraction must be chosen. For example, implementing the full complexity of the brain and the world in which it operates in a computational model would not only provide a insurmountable practical

challenge but would leave us with a problem of similar complexity to that which we face without such an implementation when attempting to understanding how such a system is able to generate the properties of cognition in which we are interested. By contrast the models used within this study function at a level of abstraction at which it is possible to maintain traction and understanding of how behaviour is supported by the underlying system as it interacts with the complex and dynamic multimodal environment.

Successful modelling also requires an interdisciplinary approach. Within this thesis models are constrained by neural and behavioural data collected across a range of populations that describe variation across both developing and mature systems. Bounding models across so many dimensions ensures a strong test of their adequacy.

Finally, modelling is an iterative process. Models are by definition not the true system and therefore operate at a level of abstraction. Their purpose is not to implement the system in its entirety but to offer a tool by which the implications of theoretical hypotheses can be tested in a principled and tractable manner. The results of which can then be used to inform the next generation of models.

Within this thesis I have examined the adequacy of interactive models of multimodal language processing and used them to generate predictions that follow from the theoretical positions they explicitly implement that can be tested empirically. Investigations have largely focused on understanding the extent to which behaviour is an emergent property of the structure of information within the complex and dynamic multimodal environment. This has necessitated a parsimonious implementation of the architecture, representations and multimodal environment involved. Given their success in capturing and offering explanation for a broad range of effects reported in the literature, they now provide valuable baseline models in which it is possible to implement additional properties of the architecture, representations or environment, in order to examine their effects in isolation and thus gain traction on complex issues regarding the nature of multimodal cognition that have occupied minds for centuries.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4), 419-439.

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, 57(4), 502-518.

Altmann, G., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33(4), 583-609.

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica*, 137(2), 181-189.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7), 280-289.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol*., 59, 617-645.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences*, 22(04), 637-660.

Berkeley, G. (1709). *An essay towards a new theory of vision*. Aaron Rhames.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6(02), 201-251.

Bruck, M., Genesee, F., & Caravolas, M. (1997). A cross-linguistic study of early literacy acquisition. *Foundations of reading acquisition and dyslexia: Implications for early intervention*, 145-162.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1), 204.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic bulletin & review*, 12(3), 453-459.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J. & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science*, 330(6009), 1359-1364.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, 25(2), 136-164.

Farnetani, E., & Recasens, D. (1997). Coarticulation and connected speech processes. *The handbook of phonetic sciences*, 371-404.

Ferreira, F., & Tanenhaus, M. K. (2007). Introduction to the special issue on language–vision interactions. *Journal of Memory and Language*, 57(4), 455-459.

Frost, R. (2012). A universal approach to modeling visual word recognition and reading: Not only possible, but also inevitable. *Behavioral and Brain Sciences*, 35(05), 310-329.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12(5-6), 613-656.

Goswami, U., Gombert, J. E., & de Barrera, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French, and Spanish. *Applied Psycholinguistics*, 19(01), 19-52.

Hanley, R., Masterson, J., Spencer, L., & Evans, D. (2004). How long do the advantages of learning to read a transparent orthography last? An investigation of the reading skills and reading impairment of Welsh children at 10 years of age. *The Quarterly Journal of Experimental Psychology: Section A*, 57(8), 1393-1410.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, 106(3), 491-528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review*, 111(3), 662-720.

Houghton, G., & Zorzi, M. (2003). Normal and impaired spelling in a connectionist dual-route architecture. *Cognitive Neuropsychology*, 20(2), 115-162.

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, 96(1), B23-B32.

Huettig, F., & Altmann, G. T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, 15(8), 985-1018.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, 57(4), 460-482.

Huettig, F., Mishra, R. K., & Olivers, C. N. (2012). Mechanisms and representations of language-mediated visual attention. *Frontiers in psychology*, 2.394.

Huettig, F., Olivers, C. N., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta psychologica*, 137(2), 138-150.

Huettig, F., Singh, N., & Mishra, R. K. (2011). Language-mediated visual orienting behavior in low and high literates. *Frontiers in psychology*, 2.

Kiyosawa, M., Itoh, M., Nakagawa, Y., Kobayashi, N., Tamai, M., 1995. Effect of kanji and kana reading on cerebral blood flow patterns measured by PET. *Jpn. J. Ophthalmol*. 39, 198–205.

Krakow, R. A. (1999). Physiological organization of syllables: a review. *Journal of Phonetics*, 27(1), 23-54.

Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and differentiation in semantic memory. *Annals of the New York Academy of Sciences*, 1124(1), 61-76.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics*, 62(3), 615-625.

Mani, N., & Huettig, F. (submitted). The changing dynamics of word-referent mapping across development.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8(1), 1-71.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, 18(1), 1-86.

McClelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive science*, 38(6), 1139-1189.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological review*, 119(4), 831.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognizeda). *The Journal of the Acoustical Society of America*, 131(1), 509-517.

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *The Quarterly Journal of Experimental Psychology*, 60(5), 661-671.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & cognition*, 37(7), 1026-1039.

Muneaux, M., & Ziegler, J. (2004). Locus of orthographic effects in spoken word recognition: Novel insights from the neighbour generation task. *Language and Cognitive Processes*, 19(5), 641-660.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review*, 115(2), 357.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23(03), 299-325.

Pattamadilok, C., Knierim, I. N., Duncan, K. J. K., & Devlin, J. T. (2010). How does learning to read affect speech perception? *The Journal of Neuroscience*, 30(25), 8435-8444.

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S. F., Cotelli, M. Cossu, G., Corte, F., Lorusso, M., Pesenti, S., Gallagher, A., Perani, D., Price, C., Frith, C. D., & Frith, U. (2000). A cultural effect on brain function. *Nature neuroscience*, 3(1), 91-96.

Perre, L., Pattamadilok, C., Montant, M., & Ziegler, J. C. (2009). Orthographic effects in spoken language: On-line activation or phonological restructuring?. *Brain research*, 1275, 73-80.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3), 280-291.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. *The emergence of language,* 381-415.

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology*, 19(7), 603-639.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science*, 38(6), 1024-1077.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological Review,* 111(1), 205-235.

Roy, D. (2005). Grounding words in perception and action: computational insights. *Trends in cognitive sciences*, 9(8), 389-396.

Rumelhart, D. E., & McClelland, J. L., & the PDP Research Group (1986). Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: Foundations & volume II: Psychological and biological models. Cambridge, MA: MIT Press.

Scharenborg, O., & Boves, L. (2010). Computational modelling of spoken-word recognition processes: Design choices and evaluation. *Pragmatics & Cognition*, 18(1), 136-164.

Seidenberg, M. S. (2013). The science of reading and its educational implications. *Language Learning and Development*, 9, 331-360.

Seymour, P. H., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94, 143-174.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558-1568.

Taft, M. (2006). Orthographically influenced abstract phonological representation: Evidence from non-rhotic speakers. *Journal of psycholinguistic research*, 35(1), 67-78.

Taft, M., & Hambly, G. (1985). The influence of orthography on phonological representations in the lexicon. *Journal of Memory and Language*, 24(3), 320-335.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, 29(6), 557-580.

Trueswell, J. C. (2008). Using eye movements as a developmental measure within psycholinguistics. Language acquisition and language disorders, 44, 73.

Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology*, 45(6), 1611-1617.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4), 625-636.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(1), 1-14.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13), 2149-2165.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin*, 131(1), 3.

Nederlandse Samenvatting

Taal verbindt informatie van verschillende *taalmodaliteiten* (spreken, luisteren, schrijven, lezen) met elkaar. Bijvoorbeeld, wanneer we de groep klanken horen die het gesproken woord "appel" vormt of de groep visuele kenmerken zien die het geschreven woord *appel* samenstellen, kunnen we snel onze kennis activeren over de vorm van een appel of over hoe en appel smaakt en aanvoelt. Daarnaast activeren we ook kennis die minder direct verwant is met de zintuiglijke ervaring van het object, zoals de informatie dat een appel gezond is en een eetbaar object is (functionele semantische kenmerken). Ook het omgekeerde is waar. Wanneer we een appel zien of de geur of smaak van een appel ervaren, kunnen we snel de geschreven of gesproken vorm van het woord appel oproepen om de naam van het object te kunnen schrijven of uitspreken.

Bovendien wordt de snelheid en het gemak waarmee ons taalverwerkingssysteem in staat is om informatie uit verschillende modaliteiten (bijv. visuele of semantische context) te activeren en te integreren duidelijk in de efficiëntie waarmee deze verschillende bronnen van informatie gebruikt kunnen worden om ambiguïteit op te lossen bij het verwerken van inherent ambigue natuurlijke taal. Dit snelle en efficiënte proces zorgt ervoor dat we ons vrijwel niet bewust zijn van de vaak grote dubbelzinnigheid van dagelijkse conversaties. Vele woorden en zinnen hebben meerdere betekenissen en ons spraaksignaal wordt in een natuurlijk setting vaak verstoord door 'ruis' (bijv. een deur die open of dicht gaat, muziek, auto's die voorbij rijden, een blaffende hond). Daarnaast veroorzaakt de spreker zelf vaak extra ambiguïteit door het maken van versprekingen, het veranderen van onderwerp, het onvolledig uitspreken van woorden, enz. Het achterhalen van de bedoelde boodschap achter een uiting als "Waar 's de bal" of zelf "Waar 's ba.." wordt echter veel eenvoudiger wanneer we de informatie uit het spraaksignaal kunnen combineren met informatie uit de directe visuele omgeving. Dit is bijvoorbeeld het geval als we de uiting horen terwijl we naar een voetbalwedstrijd kijken of op hetzelfde moment een reclameposter zien voor het afstudeerbal.

Om te begrijpen hoe mensen taal verwerken, moeten we dus de werking van de mechanismes die de interactie tussen verschillende taalmodaliteiten mogelijk maken snappen. De vraag is dus hoe en wanneer we, bij het horen van een woord of het zien van een object, kennis die gerelateerd is aan dat woord of object activeren in andere modaliteiten en in welke vorm we deze kennis activeren. Bovendien willen we weten hoe en wanneer het verwerken van een spreeksignaal of van visuele input beïnvloed wordt door informatie die al actief is in andere modaliteiten.

Om deze bredere vragen te beantwoorden, moeten we ook een antwoord vinden op de volgende vragen, bijvoorbeeld met betrekking tot wanneer informatie geactiveerd wordt: activeren we onze kennis over de vorm en/of de functie van een appel al tegen de tijd dat we "ap..." horen? En als dit zo vroeg gebeurt, activeren we dan ook visuele en/of semantische aspecten van andere woorden die ook met de klanken "ap..." beginnen, zoals 'appartement' of 'apparaat'? Verder vragen we ons af wanneer we informatie uit andere modaliteiten kunnen integreren tijdens het horen van een woord om dat woord makkelijker te kunnen begrijpen. Als we net een appel gezien zouden hebben, of als we net aan het denken waren over verhuizen naar een nieuw appartement, hoe zou dit het verwerken van de uiting "ap..." dan beïnvloeden? Bovendien, hoe zijn deze verschillende vormen van informatie met elkaar verbonden, met andere woorden, hoe is onze kennis over een woord of een voorwerp gestructureerd? Welke informatie verbindt "ap..." of "appel" met de visuele vorm van een appel? Moeten we eerst andere informatie activeren, zoals dat het een gezond stuk fruit is, om onze kennis over de vorm van een appel te kunnen activeren bij het horen van het woord "appel"? Of kunnen deze verschillende vormen van informatie tegelijk geactiveerd worden? Daarbij sluit ook de vraag aan of de verschillende vormen van informatie appart opgeslagen zijn. Wanneer we het woord 'appel' horen, kunnen we dan kennis over de betekenis van het woord activeren zonder dat we kennis over de vorm activeren en vice versa?

Deze dissertatie heeft als doel om een systeem te beschrijven dat in staat is om deze complexe en dynamische multimodale processen te ondersteunen, door middel van het aanbieden van specifieke oplossingen voor de gestelde vragen en het testen van de waarschijnlijkheid van deze oplossingen. Bovendien, aangezien ieders individuele multimodale ervaring van de wereld uniek is, wordt er in dit proefschrift onderzocht in welke mate deze unieke ervaring het multimodale systeem vormt dat multimodale taalverwerking ondersteunt. Het doel is om plausibele computationele mechanismes te definiëren die

waarschijnlijk nodig zijn en benut worden in het menselijk brein om deze multimodale componenten van menselijk taalverwerking te ondersteunen.

Aangezien veel van de gestelde vragen fundamenteel zijn voor het begrijpen van menselijke taalverwerking, is deze dissertatie natuurlijk niet de eerste waarin onderzoek is gedaan op dit gebied. Echter, door de complexiteit van de processen die ermee gemoeid zijn, zijn veel van de bovenstaande vragen nog onbeantwoord. De studies die in dit proefschrift beschreven worden, zijn bedoeld om meer grip te krijgen op zulke vragen door te concentreren op het construeren van computationele modellen (computermodellen die mentale processen simuleren) van multimodale processen die betrokken zijn bij het verwerken van afzonderlijke woorden. Daarna werd getest hoe goed de modellen in staat zijn om dezelfde patronen in gedrag en neurale activiteit te tonen als mensen tijdens het verwerken van taal in een multimodale context. Daarbij werden de modellen getest op gedrag van verschillende populaties van mensen die systematisch verschillen in structuur van de visuele en auditive relaties waaraan ze in de loop van hun leven zijn blootgesteld.

Het voordeel van het gebruiken van computationele modellen, is dat het de onderzoeker dwingt om het proces en de betrokken informatie concreet,van begin tot einde, te beschrijven. Vandaar dat een toegepast model van multimodale taalverwerking een oplossing moet bieden en dus expliciet een hypothese voor elk van de gestelde vragen moet representeren. Modellen maken het ook mogelijk om grip te krijgen op complexe en dynamische processen, zoals de processen die betrokken zijn bij het multimodaal verwerken van taal. Modellen bieden zo een manier om gevolgen van specifieke theoretische perspectieven te testen op een robuuste en gecontroleerde manier. Eenmaal een model gebouwd is, en dus ook de hypothese opgesteld is die het model toepast, kan het model getest worden. Daarbij wordt onderzocht of a) het model in staat is om de handeling in kwestie uit te voeren en b) of het model de handeling uitvoert op een manier die gelijkwaardig is aan die van een menselijk systeem (bijv. vertonen ze gedrag of interne patronen van verwerken die soortgelijk zijn aan diegene die vertoond worden door menselijke populaties). Daarna kan beoordeeld worden hoe waarschijnlijk het is dat het menselijke brein op dezelfde manier te werk gaat. Tot slot, wanneer een effectief computationeel model tot stand gebracht is, kunnen individuele eigenschappen van het model erg precies en onafhankelijk onderzocht en gemanipuleerd worden op manieren die in menselijke populaties soms niet mogelijk zijn om de processen die het model omvat beter te kunnen begrijpen.

Het modellerende kader dat gebruikt werd om de modellen in deze dissertatie the creëren, omvat cruciale eigenschappen van de neurale netwerken die deze processen toepassen in het menselijke brein. De modellen zijn samengesteld uit meerdere niet-lineaire verwerkingselementen die met elkaar verbonden zijn via *gewogen* verbindingen die samen een netwerk vormen. Wanneer een element uit het netwerk een signaal ontvangt, verzendt het een gewogen versie van dat signaal (afhankelijk van hoe sterk al de verbindingen zijn) verder naar alle andere elementen in het netwerk waar het mee verbonden is. Om de efficiëntie van het bouwen, uitvoeren en begrijpen van zulke netwerken te verbeteren, worden elementen binnen het netwerk niet gebruikt op het niveau van individuele neuronen, maar zijn ze bedoeld om het gedrag van grotere groepen van verbonden neuronen te simuleren. Wanneer netwerken gebouwd zijn, worden ze getraind om specifieke taken uit te voeren door een algoritme toe te passen dat kleine aanpassingen maakt de sterkte van de verbindingen in het netwerk. Zo leren ze vooraf gedefinieerde taken uit te voeren door te interageren met een kunstmatige leeromgeving.

Dit relatief simpele kader omvat vele complexe, maar fundamentele eigenschappen van neurale systemen, bijvoorbeeld: gedrag ontstaat door interactie met de leeromgeving; het leren is vastgelegd in de *sterkte* van de verbindingen als een vorm van langetermijngeheugen; verwerking is interactief en dynamisch en ontwikkelt in de loop van de tijd wanneer informatie doorgegeven wordt door elementen in het netwerk; representaties (bijv. van de vorm van een object of van het geluid van een gesproken woord) worden patronen van activatie verspreid over verschillende elementen in het netwerk. Al deze eigenschappen worden gevormd door de structuur van de leeromgeving.

In hoofdstuk 2 tot en met 5 van deze wordt het verklarende vermogen van het multimodale integratie model (MIM) voor taalverwerking gepresenteerd en getest. Het model biedt een compacte oplossing voor de vraag hoe modaliteiten met elkaar verbonden zijn door uit te leggen dat alle modaliteiten (enkel visuele, auditieve en functionele semantiek zijn gebruikt in het model) volledig met elkaar in verbinding staan door middel van een centrale laag in het model die alle informatie integreert (zie hoofdstuk 2, figuur 2). Dit betekent dat in zo'n systeem de beschikbare visuele of auditieve informatie meteen geïntegreerd wordt en tegelijk informatie uit andere modaliteiten kan beïnvloeden. Bijvoorbeeld het horen van "a..." kan mogelijkerwijs meteen elke kennis die geassocieerd is met deze klank (i.e. de smaak van een appel, de vorm van een appel, enz.) activeren. Deze geactiveerde kennis kan dan meteen gecombineerd worden met elke aanwezige visuele informatie om de activatie van informatie

te beperken (bijv. het horen van "a..." en het zien van een vaag bol, groen object, zou misschien meteen het activeren van kennis kunnen beïnvloeden in de richting van kennis over appels.)

Van groot belang is dat het gedrag van MIM niet op voorhand bepaald is. Het gedrag ontstaat via interactie met een kunstmatige leeromgeving die probeert om de omstandigheden waarin mensen kennis over woorden en objecten leren zo goed mogelijk na te bootsen. MIM simuleert het proces waarin personen een vorm van een object (bijv. rond, groen) leren associëren met de functie van dat object (bijv. eetbaar, voedzaam) en met de naam van het object in de vorm van een gesproken woord (bijv. "appel"). Dit werd gedaan door visuele input (e.g. een willekeurig selectie van objecten) en auditieve input (e.g. een gesproken woord dat zich ontvouwt in de tijd), te presenteren aan het netwerk. Het netwerk leerde door te proberen om de functie te activeren van het object waar het naar keek of van het object dat genoemd werd. Elke fout in deze 'voorspelling' resulteerde in kleine veranderingen in de *sterkte* van de connecties tussen elementen in het netwerk. Op die manier was er een grotere kans dat wanneer hetzelfde object getoond werd of dezelfde naam voor een object genoemd werd, het netwerk de correcte kennis zou activeren die overeenkwam met de functie van dat object. Het model gaat daarom uit van de intuïtieve hypothese dat mensen leren om informatie uit verschillende modaliteiten te associëren wanneer deze informatie herhaaldelijk tegelijkertijd in de omgeving verschijnt.

Om de validiteit van MIM als model van menselijke multimodale taalverwerking te kunnen evalueren, hebben we metingen nodig die het gedrag van de interne taalverwerking van het model kunnen vergelijken met het gedrag vertoond door mensen. *Language mediated eye gaze* (wanneer de focus van je blik beïnvloed wordt door taal) is zo'n meting. Deze methode is vaak gebruikt in psycholinguïstisch onderzoek om eigenschappen van het systeem dat achter menselijke taalverwerking zit te achterhalen. Language mediated eye gaze meet waar mensen hun blik op richten wanneer ze naar een bepaalde ‘scène’ kijken waar verschillende objecten op te zien zijn terwijl ze tegelijk gesproken zinnen horen. Deze methode is perfect geschikt voor deze studie aangezien het een gedrag (het richten van je blik op iets onder invloed van taal) is dat snel verandert, zelfs in de loop van het horen van een enkel woord en dus mogelijk de veranderingen omvat die een rol spelen in de periode waarin informatie uit visuele en auditieve verwerkingsstromen geïntegreerd wordt. Bovendien, omwille van de populariteit van de methode in de psycholinguïstiek en omdat het geen complexe expliciete reactie vergt van proefpersonen (e.g. ze worden enkel gevraagd om naar een scherm te kijken

terwijl ze luisteren naar gesproken zinnen), zijn er vele diverse data sets beschikbaar die verschillen beschrijven tussen gedrag van verschillende popluaties (e.g. jong, oud, geletterd, ongeletterd) en veranderingen in gedrag beschrijven wanneer relaties tussen auditieve en visuele verwerking systematisch gemanipuleerd worden.

Om een vergelijking mogelijk te maken tussen studies over language mediated eye gaze en het gedrag vertoond door het MIM model, creëert het model ook een output waarin het de locatie in zijn blikveld aangeeft waar het op een bepaald moment naar kijkt. Het model werd ook getraind, met behulp van hetzelfde algoritme als eerder beschreven, om naar een object te kijken wanneer de naam van dat object of een van zijn functionele eigenschappen werd geactiveerd.

Hoofdstukken 2 en 3 van deze dissertatie tonen aan dat de beschrijving van multimodale taalverwerking die gebruikt wordt in MIM in staat is om een oplossing te bieden voor een aantal language mediated eye gaze effecten die geobserveerd zijn bij geletterde volwassenen (de meest voorkomende groep proefpersonen in psycholinguïstische experimenten). Uit eerdere experimenten met deze groep proefpersonen, waarbij ze een scène met verschillende objecten zagen terwijl ze tegelijk een gesproken woord hoorden, bleek (binnen 1000 milliseconden vanaf de start van het woord) bijvoorbeeld het volgende: geletterde volwassenen kijken vaker naar objecten waarvan de naam lijkt op de beginklanken van het gesproken woord (e.g. bever, beker) dan naar objecten waarvan de naam lijkt op de eindklanken van het geproken woord (e.g. spreker, beker); kijken meer naar objecten naar mate het object sterker visueel of semantisch verwant is met het gesproken woord (bijv. gesproken woord = munt, visueel verwant object = knoop, semantisch gerelateerd object = portemonnee); kijken eerder naar objecten waarvan de naam lijkt op het begin van het gesproken woord dan naar objecten die visueel lijken op het woord of een gelijkwaardige betekenis hebben als het woord; kijken eerder naar visueel gerelateerde dan semantisch gerelateerde objecten. Het MIM model toont elk van deze eigenschappen van kijkgedrag en legt ook uit hoe deze eigenschappen het resultaat kunnen zijn van de structuur van de input of van de tijd die de informatie nodig heeft om door het netwerk te reizen dat informatie uit alle modaliteiten tegelijk integreert.

Hoofdstuk 3 onderzoekt ook waarom de eigenschappen van language mediated eye gaze veranderen in de loop van de menselijke ontwikkeling. In de loop van hun ontwikkeling wordt de blik van kinderen op objecten waarvan de naam dezelfde beginklanken heeft als een

gesproken woord (bijv. bever, beker) steeds meer doelgericht en beperkt tot de periode net nadat de beginklanken van het gesproken woord gehoord werden. Ook richt de blik van kinderen zich in de loop van de ontwikkeling steeds sneller in de richting van objecten die eigenschappen delen met de betekenis (semantiek) van het gesproken woord (bijv. gesproken woord = munt, semantische gerelateerd woord = portemonnee). Zulke data zijn eerder als bewijs aangevoerd voor een verandering in de voorkeur voor een bepaald type informatie in de loop van de ontwikkeling. Beide patronen werden vertoond door MIM gedurende zijn ontwikkeling van sterkere associaties tussen visuele en semantische eigenschappen van een object en de naam van een object. De verandering in gedrag in de loop van de ontwikkeling zou dus simpelweg het gevolg kunnen zijn van het feit dat kinderen een betere kennis van de wereld rondom hen ontwikkelen in plaats van het gevolg van een veranderende voorkeur.

Verder wordt er in hoofdstuk 3 gekeken naar de kennis die zich ontwikkelt binnen MIM. Het is denkbaar dat, aangezien het netwerk informatie uit alle modaliteiten centraal integreert, het netwerk, in de loop van het leren van verschillende multimodale associaties, zou kunnen beginnen met activeren van dezelfde kennis over een item ongeacht de modaliteit waarin het netwerk op het moment een ervaring van een representatie van het item heeft. Bijvoorbeeld, in de loop van het leerproces, leert het model dan dezelfde kennis te activeren zowel wanneer het het woord "appel" hoort als wanneer het een appel ziet? Of is de geactiveerde informatie afhankelijk van de modaliteit van de simulatie? Analyses van de kennis ontwikkeld door het model toonde dat, hoewel in het begin van de ontwikkeling kennis die geactiveerd wordt bij het zien van het object of het horen van de naam van het object meer gelijkwaardig wordt, de informatie die geactiveerd wordt via verschillende modaliteiten in latere fases van de training meer onderscheid toont. Oftewel, zelfs al hadden dezelde verwerkingselementen toegang tot informatie uit alle modaliteiten, als hun kennis over een bepaald object groter werd, gedroegen ze zich steeds meer verschillend bij blootstelling aan zijn visuele of auditieve vorm. Het model suggereert daarom dat, zelfs als er groepen neuronen in het menselijke brein zijn die informatie integreren uit meerdere modaliteiten, ze niet noodzakelijk concepten van objecten ontwikkelen die niet sterk gerelateerd zijn aan de modaliteit van zintuiglijke stimulatie.

Hoofstuk 4 rapporteert data uit een studie met language mediated eye gaze die uitgevoerd werd als deel van deze dissertatie. De studie onderzoekt de relatieve niveaus van activatie en invloed op verschillende types informatie wanneer personen taal verwerken in settings waarin op voorhand geactiveerde visuele en semantische informatie aanwezig is om het verwerken te

beïnvloeden. In twee experimenten werd de blik van de proefpersoon geregistreerd terwijl ze naar schermen keken waar meerdere objecten op afgebeeld waren terwijl ze naar gesproken woorden luisterden. De relatie tussen objecten in de geziene scène en het gehoorde woord (bijv. "cent") was systematisch gemanipuleerd. In een van de experimenten bevatten scenes telkens 1 object dat met het gesproken woord overeenkwam in alle klanken behalve de beginklank (bijv. tent). Alle andere objecten hadden geen enkele gelijkenis met het gesproken woord. In een tweede experiment was er, naast een object dat alle klanken behalve de beginklank deelde met het gesproken woord (bijv. tent), ook een object dat visueel gerelateerd was aan het gesproken woord (bijv. knoop), een object dat verwant was in betekenis (bijv. portemonnee) en een niet-gerelateerd object. In het eerste experiment keken proefpersonen vaker naar het object dat verwant was in klank met het gesproken woord dan naar de ongerelateerde objecten. Echter, wanneer proefpersonen naar scènes keken met 3 verwante objecten en 1 object dat geen gelijkenissen had met het gesproken woord, keken ze vaker naar het visueel gerelateerde en het semantisch gerelateerde object dan naar het ongerelateerde object en het object dat klanken deeldemet de naam van het object. In dit tweede experiment keken proefpersonen niet vaker naar het object dat de laatste klanken deelde dan naar het object dat geen enkele klank deelde had met het gesproken woord. Samen tonen deze data dat de visuele en semantische informatie die bij het gesproken woord hoort erg snel geactiveerd wordt, zodat latere, gedeelde klanken in het gehoorde woord geen invloed uitoefenen op het 'kijkgedrag'. Deze data suggereren dat in dagelijkse spraakverwerking, wanneer informatie in de onmiddelijke omgeving vaak aanwezig is om te helpen bij het begrijpen van de spraak, deze informatie snel aangewend en geïntegreerd wordt met het spraaksignaal. Bovendien suggereren deze data dat wanneer visuele of semantische informatie beschikbaar is binnen de directe omgeving om de interpretatie van een spraaksignaal te beïnvloeden, zulke informatie een grotere invloed kan hebben dan klanken die voorkomen aan het einde van woorden.

Het MIM model werd ook getest om te onderzoeken hoe het zich gedroeg wanneer het blootgesteld werd aan dezelfde condities als die ervaren door de proefpersonen in de twee experimenten. Het model toonde ook een vroege fixatie op visuele en semantisch gerelateerde objecten net nadat het gesproken woord begon te klinken. Bovendien, net als geobserveerd in het experiment, oefenden visuele en semantische relaties een veel grotere invloed op het fixatiegedrag dan klankgelijkenis op het einde van het gesproken woord. Dit toont verder de gelijkenis aan tussen het menselijke multimodale taalverwerkingssysteem en

een systeem waarin informatie geïntegreerd wordt uit parallelle visuele en auditieve verwerkingsstromen, zoals het systeem gebruikt in MIM.

De bevindingen die gerapporteerd worden in hoofstukken 2 tot 4 tonen aan dat MIM vaak dezelfde eigenschappen van language mediated eye gaze vertoond als geletterde volwassenen en ook een verklaring biedt voor deze eigenschappen. Deze hoofdstukken betogen dus dat het waarschijnlijk is dat MIM belangrijke eigenschappen omvat van het multimodale systeem dat dit gedrag ondersteunt in deze populatie. Een recente studie over language mediated eye gaze toont echter aan dat volwassenen die niet hebben leren lezen of schrijven verschillen vertonen in hun language mediated gaze ten opzichte van volwassenen die wel geletterd zijn. In hoofdstuk 5 gebruiken we MIM om te testen welke aspecten van het onderliggende multimodale taalverwerkingssysteem zouden kunnen verschillen in deze twee populaties die het verschil in language mediated eye gaze gedrag zouden kunnen verklaren. Binnen de studie zagen een groep geletterde en een groep ongeletterde volwassenen scènes met verschillende objecten terwijl ze een gesproken woord hoorden (bijv. beker) terwijl geregistreerd werd waar ze hun blik op richtten. Binnen de scènes waren er objecten die ongerelateerd waren aan het gesproken woord, objecten die de eerste klanken van hun naam deelden met het woord (bever), en objecten die verwant waren aan het woord in een bepaald aspect van hun betekenis (vork). Wanneer geletterde volwassenen het woord hoorden, keken ze snel naar het object met een naam die de eerste klanken deelde met het woord (bijv. be…). Daarna wendden ze snel de focus van hun blik af van het object vanaf het moment dat de klanken niet meer gedeeld werden (..ker) en keken in de plaats daarvan meer naar het object dat verwant was in betekenis. Ongeletterde volwassenen daarentegen keken enkel meer naar het object dat verwant was in betekenis (in een experiment met identieke voorwerpen en woorden). Ze keken even vaak naar objecten die de eerste klanken deelden met het woord als naar objecten die geen relatie hadden met het gesproken woord. Een tweede experiment toonde aan dat analfabeten wel vaker naar klankgerelateerde objecten keken wanneer zo'n object gepresenteerd werd in een scène met verder alleen maar ongerelateerde objecten. Toch bleef het verschil klein en de focus van hun blik was niet gelinkt aan de korte tijd waarin klanken van het woord en het object gedeeld werden, zoals wel het geval was voor geletterde volwassenen.

Eerder onderzoek naar de effecten van geletterdheid op taalverwerking suggereerde dat leren lezen en schrijven veranderingen teweeg brengt in de manier waarop mensen de klanken die een woord vormen representeren. Sommige onderzoekers suggereerden dat leren lezen en

schrijven ons meer bewust maakt van de individuele klanken waaruit een woord bestaat en dat als we niet lezen en schrijven, we een woord eerder representeren op het niveau van groepen klanken (e.g. op het niveau van lettergrepen of zelfs woorden). Om te testen of zulke veranderingen als het gevolg van geletterdheid het verschil kunnen verklaren tussen de language mediated eye gaze van geletterden en ongeletterden, werd het ongeletterde systeem gesimuleerd binnen MIM door de structuur van de informatie die geactiveerd wordt door het horen van een gesproken woord te manipuleren. In de loop van het horen van het gesproken woord, activeert het ongeletterde MIM systeem kennis van de klanken die het woord vormen. Echter, deze kennis kan enkel onderscheid maken tussen klanken op het niveau van individuele lettergrepen of woorden. Met deze manipulatie, genereerde het MIM systeem kijkgedrag dat soortgelijk was aan het gedrag van ongeletterden. Dit suggereert dat het verschil tussen geletterden en ongeletterden in hun kennis van de klankstructuur van woorden, het verschil dat ze tonen in kijkgedrag kan verklaren.

In hoofdstuk 6 van deze dissertatie is er verder onderzocht hoe leren lezen en schrijven tot verschillen in populaties kan leiden wat betreft hun multimodale taalverwerking. Leren lezen vergt van een persoon om zijn of haar kennis van een taal aan te spreken via een een systeem van visuele symbolen (i.e. het schrijfsysteem). In dit hoofdstuk concentreren we op verschillen tussen geletterde populaties die kunnen ontstaan uit verschillen in de structuur van het schrijfsysteem waarin ze geletterd zijn.

Schrijfsystemen verschillen erg in hoe ze betekenis en klankstructuren van woorden 'coderen'. In het alfabetische schrijfsystemen zoals Nederlands of Engels, worden individuele letters of groepen letters gebruikt om individuele klanken te respresenteren die gecombineerd kunnen worden om de woorden van de taal te representeren (hoewel de samenhang van deze relaties verschilt tussen alfabetische systemen, bijv. de samenhang tussen letters en klanken is groter in het Nederlands dan in het Engels. Logografische schrijfsystemen zoals Chinees daarentegen coderen de klankstructuur van een taal niet veel verder dan het niveau van individuele woorden . In andere woorden, de geschreven vorm van een Chinees woord levert erg weinig, of geen, informatie over de individuele klanken die samen het woord vormen. Toch, anders dan alfabetische systemen, kunnen in het Chinees sommige componenten van het schrijfsysteem informatie geven over de betekenis van het woord. Dit niveau van variatie tussen schrijfsystemen heeft tot de suggestie geleid dat verschillende schrijfsystemen waarschijnlijk andere compationele mechanismes nodig hebben om het lezen te ondersteunen.

Binnen hoofdstuk 6 is er aangetoond, door toepassingen in computationele modellen, dat één enkel computationeel ontwerp het lezen kan ondersteunen voor alle schrijfsystemen in de wereld. Door een aantal verschillende, uiteenlopende schrijfsystemen te implementeren in hetzelfde ontwerp, maar in verschillende netwerken, onderzoek ik hoe verschillen in hoe schrijfsystemen de klank en betekenis van woorden coderen het leren lezen beïnvloedt en hoe kennis van de taal opgeslagen en geraadpleegd wordt. De resultaten tonen bijvoorbeeld aan dat regelmatige overeenkomst tussen de basiselementen van het schrijfsysteem en de klankstructuur van een klank en/of het betekenissysteem van een taal resulteert in netwerken die minder training nodig heben om de klank en de betekenis van het geschreven woorden boven te halen. Bovendien tonen simulaties aan dat het manipuleren van deze overeenkomst tussen de geschreven componenten en de klank/betekenis die ze representeren, de manier waarop activatie verspreid wordt over neurale ‘paden’ in het netwerk verandert. Dit komt overeen met neuro-imaging studies die soortgelijke verschillen aantoonden in de neurale paden in het menselijke brein die gebruikt worden om te lezen tussen populaties die verschillen in het schrijfsysteem waarin ze getraind waren. Tot slot, door het vergelijken van de structuur van de kennis die opgeslagen is in netwerken die getraind zijn op verschillende schrijfsystemen, toon ik aan dat deze eigenschappen van het schrijfsysteem de aard van de opgeslagen structuur van betekenis en klank kan beïnvloeden.

Samengevat, in deze dissertatie bied ik een omschrijving van de mechanismes die waarschijnlijk de menselijke multimodale taalverwerking ondersteunen, door gebruik te maken van computationele modellen. Ik toon verder de aanneembaarheid van de gepresenteerde modellen aan, door te laten zien dat ze in staat zijn om soortgelijk gedrag te vertonen als de mens, en neurale processen weet na te bootsen als mensen een taak uitvoeren waar intergratie van verschillende modaliteiten centraal staat.

Acknowledgements

I must first say a huge “Dank je wel” and “Danke schön” to the Dutch and German public who have funded this thesis. Especially the residents of Nijmegen who have made my partner and I feel so welcome here over the past four years and been so forgiving of our woeful knowledge of the Dutch language (we are still working on it!).

I also wish to acknowledge the role of the reading committee (Prof. Ardi Roelofs, Prof. Chris Olivers and Dr. Stefan Frank) and thank them for dedicating the substantial amount of time required to examine this thesis, especially Dr. Stefan Frank for his valuable comments and feedback.

This thesis has been facilitated and shaped by many individuals I have been fortunate to have known as my colleagues, friends and family and I wish to thank you all even though it is not possible for me to mention you all in person.

To my colleagues at the MPI, thank you for making this institute such a unique, stimulating environment in which to study language processing. The rigour, dedication and profound knowledge of the discipline that you display daily in your work is an inspiration that drives us all forward. I have benefitted greatly from being surrounded by such sharp and hardworking individuals. I must give special mention to Karin Kastens and Meggie Uijen who ensured the library, which at times became my second home, remained a peaceful and productive place of work.

To my PoL colleagues thank you for engaging with my studies even if they were often only loosely related to your own. Thanks also to Annelies van Wijngaarden for allowing me to make use of your voice in my experiments and Caitlin Decuyper and Suzanne Jongman for translating the summary of this thesis into Dutch. A special thank you also to Florian Hintz, Joost Rommers and Agnieska Knopoka who always provided insightful comments and questions.

To my office mate Florian Hintz, I feel incredibly fortunate to have had you as my office mate through the most crucial stages of my PhD. Thank you for keeping me in the loop both socially and on many occasions academically. For never ceasing to provide a friendly face and atmosphere, for entertaining my frequent ramblings and laughing at my very bad jokes. It was a pleasure to share the highs and lows experienced over the past three years with you, your friendship throughout kept me feeling informed, included and sane (most of the time).

To my supervisory team Antje Meyer, Falk Huettig and Padraic Monaghan, I greatly appreciate the courage and trust you displayed throughout this project. Antje and Falk, embarking on such a project reliant on a methodology and student new to you both must have been daunting, but you maintained belief and offered the freedom necessary for the project to succeed. Throughout you consistently offered valuable guidance in addition to insightful and piercing questions in areas often outside your own area of expertise. Antje, my partner and I thank you for the great compassion and understanding you displayed during the most difficult of periods. You lifted what could have been an enormous weight at a very difficult time.

Falk, throughout the four years you have never wavered from offering a warm, friendly and positive voice even in what must have been the most infuriating moments. You were always available to provide time and hope whenever required. Your management of my erratic productivity has been exemplary, providing the necessary carrots and sticks whenever they were necessary. You have continually displayed enthusiasm and boldness in driving this project forwards positively, always aiming high, never shying away from a challenge or entertaining negativity. I have learnt immensely from your efficient, bold and positive approach to conducting research.

I also wish to thank my sister Heidi for her brilliant translation of an incredibly vague waffling description into my fantastic cover image. I also thank Heidi and my brothers Harvey and Laurence not only for their love but also for the work ethic, perseverance and self-discipline they demonstrate. Their example has driven this thesis forward on countless occasions.

Finally, there are four individuals that have not only had a profound influence on the completion of this PhD but also on the course of my life. I will therefore now embark on the final insurmountable challenge of this thesis, trying to communicate my gratitude (without too many clichés) to those that have gone above and beyond in getting me to and through to the completion of this PhD.

To Padraic, I will always be grateful for the immense levels of belief, patience and generosity you have displayed since I walked into your office as a bundle of nerves nearly 7 years ago. Throughout this period you have remained the model tutor, mentor, academic and human being. I have always felt you have had my best interests in mind, offering me every possible opportunity you've been able to provide for me to develop and succeed, even when this may have clashed with your own interests. You have always created time and never implied the burden that this must have presented, whether it was the weekly skype chats in the first year of this PhD or the hours you must have dedicated in the weekends and evenings sifting through my ramblings to extract and formulate a coherent message. You have remained throughout irrepressibly positive, insightful and genuinely sensitive to the emotional journey of the student. I have learnt so much about so much, it has been a great pleasure and privilege to have been your student for the past 7 years.

To my parents Josephine and Brian, none of this would have been possible without the selflessness you have displayed throughout my life. You provided every possible opportunity and freedom you were able to offer, so that I was able to pursue my interests and search for a fulfilling career, while also providing a strong, warm and unconditionally loving home to fall back on whenever it was needed. I thank you for instilling within me an appreciation for the value of knowledge and the pleasures its pursuit can bring in addition to ensuring I am well-schooled in the art of objective positivity.

And so finally, to my loving partner Laura. I apologise for the numerous lonely evenings and weekends you have had to endure over the past four years. You have turned these incredibly challenging years into the most wonderful of journeys. Yet the challenges I faced were nothing compared to the challenges you have overcome and the personal sacrifices you have made to make completion of this PhD possible. Throughout you displayed immense trust, never questioning my decisions to dedicate so much time to something that must seem so remote and obscure. Your unconditional love and support has never wavered, you provided the constant, the rejuvenating life away from work that has ensured these years were so immensely rewarding and enjoyable.

Thank you.

Publications

Smith, A. C., Monaghan, P., & Huettig, F. (in preparation). The effects of orthographic transparency on the reading system: Insights from a computational model of reading development.

Smith, A. C., Monaghan, P., & Huettig, F. (in preparation). Connecting language and vision: A multimodal integration model (MIM) of language mediated visual attention.

Smith, A. C., Monaghan, P., & Huettig, F. (submitted). The multimodal nature of spoken word processing in the visual world: Testing the predictions of a Multimodal Integration Model (MIM).

Smith, A. C., Monaghan, P., & Huettig, F. (submitted). Complex word recognition behaviour emerges from the richness of the word learning environment.

Monaghan, P., Mattock, K., Davies, R., & Smith, A. (2015). Gavagai is as gavagai does: Learning nouns and verbs from cross-situational statistics. *Cognitive Science*, 39, 1099-1112.

Smith, A. C., Monaghan, P., & Huettig, F. (2014). Literacy effects on language and vision: Emergent effects from an amodal shared resource (ASR) computational model. *Cognitive Psychology*. 75, 28-54.

Smith, A. C., Monaghan, P., & Huettig, F. (2014). A comprehensive model of spoken word recognition must be multimodal: Evidence from studies of language mediated visual attention. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Smith, A. C., Monaghan, P., & Huettig, F. (2014). Examining strains and symptoms of the 'Literacy Virus': The effects of orthographic transparency on phonological processing in a connectionist model of reading. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds*.), Proceedings of the 36th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Smith, A. C., Monaghan, P., & Huettig, F. (2014). Modelling language-vision interactions in the hub-and-spoke framework. In J. Mayor, & P. Gomez (Eds.), *Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop (NCPW13).* Singapore: World Scientific Publishing.

Smith, A. C., Monaghan, P., & Huettig, F. (2013). An amodal shared resource model of language-mediated visual attention. *Frontiers in Psychology*. 4:528.

Smith, A. C., Monaghan, P., & Huettig, F. (2013). Modelling the effects of formal literacy training on language-mediated visual attention. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 3420-3425). Austin, TX: Cognitive Science Society.

Smith, A., & Monaghan, P. (2011). What are the functional units in reading? Evidence for statistical variation influencing word processing. In E. J. Davelaar (Eds.), *Connectionist Models of Neurocognition and Emergent Behavior from Theory to Applications: Proceedings of the 12th Neural Computation and Psychology Workshop (NCPW12).* (pp. 159-172). Singapore: World Scientific Publishing.

Curriculum Vitae

Alastair Charles Smith (Birmingham, United Kingdom, 1983) obtained his bachelor's degree in Combined Science (Mathematics, Computer Science & Psychology, BSc) from Lancaster University, United Kingdom in 2010. This was followed by a master's degree in Cognition & Computational Neuroscience (M.Res.) from the University of Birmingham, United Kingdom in 2011. He began his PhD project in the same year at the Max Planck Institute for Psycholinguistics, Nijmegen. The results of his PhD research, conducted in the Psychology of Language Department, are detailed in this thesis. He is currently a postdoctoral researcher within the department.

MPI Series in Psycholinguistics

1. *The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing.* Miranda van Turennout
2. *The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography.* Niels O. Schiller
3. *Lexical access in the production of ellipsis and pronouns.* Bernadette M. Schmitt
4. *The open-/closed-class distinction in spoken-word recognition.* Alette Haveman
5. *The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach.* Kay Behnke
6. *Gesture and speech production.* Jan-Peter de Ruiter
7. *Comparative intonational phonology: English and German. Esther* Grabe
8. *Finiteness in adult and child German.* Ingeborg Lasser
9. *Language input for word discovery.* Joost van de Weijer
10. *Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe.* James Essegbey
11. *Producing past and plural inflections.* Dirk Janssen
12. *Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea.* Anna Margetts
13. *From speech to words.* Arie van der Lugt
14. *Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language.* Eva Schultze-Berndt
15. *Interpreting indefinites: An experimental study of children's language comprehension.* Irene Krämer
16. *Language-specific listening: The case of phonetic sequences.* Andrea Weber
17. *Moving eyes and naming objects.* Femke van der Meulen

18. *Analogy in morphology: The selection of linking elements in Dutch compounds.* Andrea Krott

19. *Morphology in speech comprehension.* Kerstin Mauth

20. *Morphological families in the mental lexicon.* Nivja H. de Jong

21. *Fixed expressions and the production of idioms.* Simone A. Sprenger

22. *The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria).* Birgit Hellwig

23. *Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies.* Fermín Moscoso del Prado Martín

24. *Contextual influences on spoken-word processing: An electrophysiological approach.* Daniëlle van den Brink

25. *Perceptual relevance of prevoicing in Dutch.* Petra M. van Alphen

26. *Syllables in speech production: Effects of syllable preparation and syllable frequency.* Joana Cholin

27. *Producing complex spoken numerals for time and space.* Marjolein Meeuwissen

28. *Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction.* Rachèl J. J. K. Kemps

29. *At the same time...: The expression of simultaneity in learner varieties.* Barbara Schmiedtová

30. *A grammar of Jalonke argument structure.* Friederike Lüpke

31. *Agrammatic comprehension: An electrophysiological approach.* Marlies Wassenaar

32. *The structure and use of shape-based noun classes in Miraña (North West Amazon).* Frank Seifart

33. *Prosodically-conditioned detail in the recognition of spoken words.* Anne Pier Salverda

34. *Phonetic and lexical processing in a second language.* Mirjam Broersma

35. *Retrieving semantic and syntactic word properties.* Oliver Müller

36. *Lexically-guided perceptual learning in speech processing.* Frank Eisner

37. *Sensitivity to detailed acoustic information in word recognition.* Keren B. Shatzman

38. *The relationship between spoken word production and comprehension.* Rebecca Özdemir

39. *Disfluency: Interrupting speech and gesture.* Mandana Seyfeddinipur
40. *The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation.* Christiane Dietrich
41. *Cognitive cladistics and the relativity of spatial cognition.* Daniel B.M. Haun
42. *The acquisition of auditory categories.* Martijn Goudbeek
43. *Affix reduction in spoken Dutch.* Mark Pluymaekers
44. *Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence.* Valesca Kooijman
45. *Space and iconicity in German Sign Language (DGS).* Pamela Perniss
46. *On the production of morphologically complex words with special attention to effects of frequency.* Heidrun Bien
47. *Crosslinguistic influence in first and second languages: Convergence in speech and gesture.* Amanda Brown
48. *The acquisition of verb compounding in Mandarin Chinese.* Jidong Chen
49. *Phoneme inventories and patterns of speech sound perception.* Anita Wagner
50. *Lexical processing of morphologically complex words: An information-theoretical perspective.* Victor Kuperman
51. *A grammar of Savosavo, a Papuan language of the Solomon Islands.* Claudia Wegener
52. *Prosodic structure in speech production and perception.* Claudia Kuzla
53. *The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension.* Sarah Schimke
54. *Studies on intonation and information structure in child and adult German.* Laura de Ruiter
55. *Processing the fine temporal structure of spoken words.* Eva Reinisch
56. *Semantics and (ir)regular inflection in morphological processing.* Wieke Tabak
57. *Processing strongly reduced forms in casual speech.* Susanne Brouwer
58. *Ambiguous pronoun resolution in L1 and L2 German and Dutch.* Miriam Ellert
59. *Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging.* Ian FitzPatrick

60. *Processing casual speech in native and non-native language.* Annelie Tuinman
61. *Split intransitivity in Rotokas, a Papuan language of Bougainville.* Stuart Robinson
62. *Evidentiality and intersubjectivity in Yurakaré: An interactional account.* Sonja Gipper
63. *The influence of information structure on language comprehension: A neurocognitive perspective.* Lin Wang
64. *The meaning and use of ideophones in Siwu.* Mark Dingemanse
65. *The role of acoustic detail and context in the comprehension of reduced pronunciation variants.* Marco van de Ven
66. *Speech reduction in spontaneous French and Spanish.* Francisco Torreira
67. *The relevance of early word recognition: Insights from the infant brain.* Caroline Junge
68. *Adjusting to different speakers: Extrinsic normalization in vowel perception.* Matthias J. Sjerps
69. *Structuring language. Contributions to the neurocognition of syntax.* Katrien R. Segaert
70. *Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading.* Birgit Knudsen
71. *Gaze behavior in face-to-face interaction.* Federico Rossano
72. *Sign-spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space.* Conny de Vos
73. *Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning.* Attila Andics
74. *Lexical processing of foreign-accented speech: Rapid and flexible adaptation.* Marijt Witteman
75. *The use of deictic versus representational gestures in infancy.* Daniel Puccini
76. *Territories of knowledge in Japanese conversation.* Kaoru Hayano
77. *Family and neighbourhood relations in the mental lexicon: A cross-language perspective.* Kimberley Mulder
78. *Contributions of executive control to individual differences in word production.* Zeshu Shao
79. *Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing.* Patrick van der Zande

80. *High pitches and thick voices: The role of language in space-pitch associations.* Sarah Dolscheid

81. *Seeing what's next: Processing and anticipating language referring to objects.* Joost Rommers

82. *Mental representation and processing of reduced words in casual speech.* Iris Hanique

83. *The many ways listeners adapt to reductions in casual speech.* Katja Poellmann

84. *Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners.* Giuseppina Turco

85. *Morphological processing in younger and older people: Evidence for flexible dual-route access.* Jana Reifegerste

86. *Semantic and syntactic constraints on the production of subject-verb agreement.* Alma Veenstra

87. *The acquisition of morphophonological alternations across languages.* Helen Buckler

88. *The evolutionary dynamics of motion event encoding.* Annemarie Verkerk

89. *Rediscovering a forgotten language.* Jiyoun Choi

90. *The road to native listening: Language-general perception, language-specific input.* Sho Tsuji

91. *Infants'understanding of communication as participants and observers.* Gudmundur Bjarki Thorgrímsson

92. *Information structure in Avatime.* Saskia van Putten

93. *Switch reference in Whitesands.* Jeremy Hammond

94. *Machine learning for gesture recognition from videos.* Binyam Gebrekidan Gebre

95. *Acquisition of spatial language by signing and speaking children: a comparison of Turkish sign language (TID) and Turkish.* Beyza Sümer

96. *An ear for pitch: on the effects of experience and aptitude in processing pitch in language and music.* Salomi Savvatia Asaridou

97. *lncrementality and Flexibility in Sentence Production.* Maartje van de Velde

98. *Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture.* Edwin van Leeuwen

99. *The request system in Italian interaction.* Giovanni Rossi

100. *Timing turns in conversation: A temporal preparation account.* Lilla Magyari

101. *Assessing birth language memory in young adoptees.* Wencui Zhou
102. *A social and neurobiological approach to pointing in speech and gesture.* David Peeters
103. *Investigating the genetic basis of reading and language skills.* Alessandro Gialluisi
104. *Conversation electrified: The electrophysiology of spoken speech act recognition.* Rósa Signý Gisladottir
105. *Modelling multimodal language processing.* Alastair Charles Smith